From Averaging to Acceleration, There is Only a Step-size

Nicolas Flammarion, Francis Bach

From Averaging to Acceleration, There is Only a Step-size

Nicolas Flammarion and Francis Bach
INRIA - Sierra project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France
nicolas.flammarion@ens.fr, francis.bach@ens.fr

April 7, 2015

Abstract

We show that accelerated gradient descent, averaged gradient descent and the heavy-ball method for non-strongly-convex problems may be reformulated as constant parameter second-order difference equation algorithms, where stability of the system is equivalent to convergence at rate $O(1/n^2)$, where n is the number of iterations. We provide a detailed analysis of the eigenvalues of the corresponding linear dynamical system, showing various oscillatory and non-oscillatory behaviors, together with a sharp stability result with explicit constants. We also consider the situation where noisy gradients are available, where we extend our general convergence result, which suggests an alternative algorithm (i.e., with different step sizes) that exhibits the good aspects of both averaging and acceleration.

1 Introduction

Many problems in machine learning are naturally cast as convex optimization problems over a Euclidean space; for supervised learning this includes least-squares regression, logistic regression, and the support vector machine. Faced with large amounts of data, practitioners often favor first-order techniques based on gradient descent, leading to algorithms with many cheap iterations. For smooth problems, two extensions of gradient descent have had important theoretical and practical impacts: acceleration and averaging.

Acceleration techniques date back to Nesterov (1983) and have their roots in momentum techniques and conjugate gradient (Polyak, 1987). For convex problems, with an appropriately weighted momentum term which requires to store two iterates, Nesterov (1983) showed that the traditional convergence rate of $O(1/n)$ for the function values after n iterations of gradient descent goes down to $O(1/n^2)$ for accelerated gradient descent, such a rate being optimal among first-order techniques that can access only sequences of gradients (Nesterov, 2004). Like conjugate gradient methods for solving linear systems, these methods are however more sensitive to noise in the gradients; that is, to preserve their improved convergence rates, significantly less noise may be tolerated (d'Aspremont, 2008; Schmidt et al., 2011; Devolder et al., 2014).

Averaging techniques which consist in replacing the iterates by the average of all iterates have also been thoroughly considered, either because they sometimes lead to simpler proofs, or because they lead to improved behavior. In the noiseless case where gradients are exactly available, they do not improve the convergence rate in the convex case; worse, for strongly-convex problems, they are not linearly convergent while regular gradient descent is. Their main advantage comes with random unbiased gradients, where it has been shown that they lead to better convergence rates than the unaveraged counterparts, in particular because they allow larger step-sizes (Polyak and Juditsky, 1992; Bach and Moulines, 2011). For example, for least-squares regression with stochastic gradients, they lead to convergence rates of $O(1/n)$, even in the non-strongly convex case (Bach and Moulines, 2013).

In this paper, we show that for quadratic problems, both averaging and acceleration are two instances of the same second-order finite difference equation, with different step-sizes. They may thus be analyzed jointly, together with a non-strongly convex version of the heavy-ball method (Polyak, 1987, Section 3.2). In presence of random zero-mean noise on the gradients, this joint analysis allows to design a novel intermediate algorithm that exhibits the good aspects of both acceleration (quick forgetting of initial conditions) and averaging (robustness to noise).

In this paper, we make the following contributions:

- We show in Section 2 that accelerated gradient descent, averaged gradient descent and the heavy-ball method for non-strongly-convex problems may be reformulated as constant parameter second-order difference equation algorithms, where stability of the system is equivalent to convergence at rate $O(1/n^2)$.
- In Section 3, we provide a detailed analysis of the eigenvalues of the corresponding linear dynamical system, showing various oscillatory and non-oscillatory behaviors, together with a sharp stability result with explicit constants.
- In Section 4, we consider the situation where noisy gradients are available, where we extend our general convergence result, which suggests an alternative algorithm (i.e., with different step sizes) that exhibits the good aspects of both averaging and acceleration.
- In Section 5, we illustrate our results with simulations on synthetic examples.

2 Second-Order Iterative Algorithms for Quadratic Functions

Throughout this paper, we consider minimizing a convex quadratic function $f : \mathbb{R}^d \to \mathbb{R}$ defined as:

$$f(\theta) = \frac{1}{2}\langle \theta, H\theta \rangle - \langle q, \theta \rangle, \tag{1}$$

with $H \in \mathbb{R}^{d \times d}$ a symmetric positive semi-definite matrix and $q \in \mathbb{R}^d$. Without loss of generality, H is assumed invertible (by projecting onto the orthogonal of its null space), though its eigenvalues could be arbitrarily small. The solution is known to be $\theta_* = H^{-1}q$,

but the inverse of the Hessian is often too expensive to compute when d is large. The excess cost function may be simply expressed as $f(\theta_n) - f(\theta_*) = \frac{1}{2}\langle \theta_n - \theta_*, H(\theta_n - \theta_*)\rangle$.

2.1 Second-order algorithms

In this paper we study second-order iterative algorithms of the form:

$$\theta_{n+1} = A_n\theta_n + B_n\theta_{n-1} + c_n, \tag{2}$$

started with $\theta_1 = \theta_0$ in $\mathbb{R}^d$, with $A_n \in \mathbb{R}^{d\times d}$, $B_n \in \mathbb{R}^{d\times d}$ and $c_n \in \mathbb{R}^d$ for all $n \in \mathbb{N}^*$. We impose the natural restriction that the optimum θ_* is a stationary point of this recursion, that is, for all $n \in \mathbb{N}^*$:

$$\theta_* = A_n\theta_* + B_n\theta_* + c_n. \qquad (\theta_*\text{-stationarity})$$

By letting $\phi_n = \theta_n - \theta_*$ we then have $\phi_{n+1} = A_n\phi_n + B_n\phi_{n-1}$, started from $\phi_0 = \phi_1 = \theta_0 - \theta_*$. Thus, we restrict our problem to the study of the convergence of an iterative system to 0.

In connection with accelerated methods, we are interested in algorithms for which $f(\theta_n) - f(\theta_*) = \frac{1}{2}\langle \phi_n, H\phi_n\rangle$ converges to 0 at a speed of $O\left(1/n^2\right)$. Within this context we impose that A_n and B_n have the form :

$$A_n = \frac{n}{n+1}A \text{ and } B_n = \frac{n-1}{n+1}B \quad \forall n \in \mathbb{N} \text{ with } A, B \in \mathbb{R}^{d\times d}. \qquad (\text{n-scalability})$$

By letting $\eta_n = n\phi_n = n(\theta_n - \theta_*)$, we can now study the simple iterative system with *constant* terms $\eta_{n+1} = A\eta_n + B\eta_{n-1}$, started at $\eta_0 = 0$ and $\eta_1 = \theta_0 - \theta_*$. Showing that the function values remain bounded, we directly have the convergence of $f(\theta_n)$ to $f(\theta_*)$ at the speed $O\left(1/n^2\right)$. Thus the n-scalability property allows to switch from a convergence problem to a stability problem.

For feasibility concerns the method can only access H through matrix-vector products. Therefore A and B should be polynomials in H and c a polynomial in H times q, if possible of low degree. The following theorem clarifies the general form of iterative systems which share these three properties (see proof in Appendix B).

Theorem 1. *Let $(P_n, Q_n, R_n) \in (\mathbb{R}[X])^3$ for all $n \in \mathbb{N}$, be a sequence of polynomials. If the iterative algorithm defined by Eq. (2) with $A_n = P_n(H)$, $B_n = Q_n(H)$ and $c_n = R(H)q$ satisfies the θ_*-stationarity and n-scalability properties, there are polynomials $(\bar{A}, \bar{B}) \in (\mathbb{R}[X])^2$ such that:*

$$\begin{aligned} A_n &= 2\frac{n}{n+1}\left(I - \left(\frac{\bar{A}(H) + \bar{B}(H)}{2}\right)H\right), \\ B_n &= -\frac{n-1}{n+1}\left(I - \bar{B}(H)H\right) \quad \text{and} \quad c_n = \left(\frac{n\bar{A}(H) + \bar{B}(H)}{n+1}\right)q. \end{aligned}$$

Note that our result prevents A_n and B_n from being zero, thus requiring the algorithm to strictly be of second order. This illustrates the fact that first-order algorithms as gradient descent do not have the convergence rate in $O(1/n^2)$.

We now restrict our class of algorithms to lowest possible order polynomials, that is, $\bar{A} = \alpha I$ and $\bar{B} = \beta I$ with $(\alpha, \beta) \in \mathbb{R}^2$, which correspond to the fewest matrix-vector products per iteration, leading to the *constant-coefficient* recursion for $\eta_n = n\phi_n = n(\theta_n - \theta_*)$:

$$\eta_{n+1} = (I - \alpha H)\,\eta_n + (I - \beta H)\,(\eta_n - \eta_{n-1})\,. \tag{3}$$

Expression with gradients of f. The recursion in Eq. (3) may be written with gradients of f in multiple ways. In order to preserve the parallel with accelerated techniques, we rewrite it as:

$$\theta_{n+1} = \frac{2n}{n+1}\theta_n - \frac{n-1}{n+1}\theta_{n-1} - \frac{n\alpha+\beta}{n+1} f'\left(\frac{n(\alpha+\beta)}{n\alpha+\beta}\theta_n - \frac{(n-1)\beta}{n\alpha+\beta}\theta_{n-1}\right). \tag{4}$$

It may be interpreted as a modified gradient recursion with two potentially different affine (i.e., with coefficients that sum to one) combinations of the two past iterates. This reformulation will also be crucial when using noisy gradients. The allowed values for $(\alpha, \beta) \in \mathbb{R}^2$ will be determined in the following sections.

2.2 Examples

Averaged gradient descent. We consider averaged gradient descent (referred to from now on as "Av-GD") (Polyak and Juditsky, 1992) with step-size $\gamma \in \mathbb{R}$ defined by:

$$\psi_{n+1} = \psi_n - \gamma f'(\psi_n), \qquad \theta_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} \psi_i.$$

When computing the average online as $\theta_{n+1} = \theta_n + \frac{1}{n+1}(\psi_{n+1} - \theta_n)$ and seeing the average as the main iterate, the algorithm becomes (see proof in Appendix B.2):

$$\theta_{n+1} = \frac{2n}{n+1}\theta_n - \frac{n-1}{n+1}\theta_{n-1} - \frac{\gamma}{n+1} f'\big(n\theta_n - (n-1)\theta_{n-1}\big).$$

This corresponds to Eq. (4) with $\alpha = 0$ and $\beta = \gamma$.

Accelerated gradient descent. We consider the accelerated gradient descent (referred to from now on as "Acc-GD") (Nesterov, 1983) with step-sizes $(\gamma, \delta_n) \in \mathbb{R}^2$:

$$\theta_{n+1} = \omega_n - \gamma f'(\omega_n), \qquad \omega_n = \theta_n + \delta_n(\theta_n - \theta_{n-1}).$$

For smooth optimization the accelerated literature (Nesterov, 2004; Beck and Teboulle, 2009) uses the step-size $\delta_n = 1 - \frac{3}{n+1}$ and their results are not valid for bigger step-size δ_n. However $\delta_n = 1 - \frac{2}{n+1}$ is compatible with the framework of Lan (2012) and is more convenient for our set-up. This corresponds to Eq. (4) with $\alpha = \gamma$ and $\beta = \gamma$. Note that accelerated techniques are more generally applicable, e.g., to composite optimization with smooth functions (Nesterov, 2013; Beck and Teboulle, 2009).

Heavy ball. We consider the heavy-ball algorithm (referred to from now on as "HB") (Polyak, 1964) with step-sizes $(\gamma, \delta_n) \in \mathbb{R}^2$:

$$\theta_{n+1} = \theta_n - \gamma f'(\theta_n) + \delta_n(\theta_n - \theta_{n-1}),$$

when $\delta_n = 1 - \frac{2}{n+1}$. We note that typically δ_n is constant for strongly-convex problems. This corresponds to Eq. (4) with $\alpha = \gamma$ and $\beta = 0$.

3 Convergence with Noiseless Gradients

We study the convergence of the iterates defined by: $\eta_{n+1} = (I - \alpha H)\,\eta_n + (I - \beta H)\,(\eta_n - \eta_{n-1})$. This is a second-order iterative system with constant coefficients that it is standard to cast in a linear framework (see, e.g., Ortega and Rheinboldt, 2000). We may rewrite it as:

$$\Theta_n = F\Theta_{n-1}, \quad \text{with } \Theta_n = \begin{pmatrix} \eta_n \\ \eta_{n-1} \end{pmatrix} \text{ and } F = \begin{pmatrix} 2I - (\alpha+\beta)\,H & \beta H - I \\ I & 0 \end{pmatrix} \in \mathbb{R}^{2d \times 2d}.$$

Thus $\Theta_n = F^n \Theta_0$. Following O'Donoghue and Candes (2013), if we consider an eigenvalue decomposition of H, i.e., $H = P\mathrm{Diag}(h)P^\top$ with P an orthogonal matrix and (h_i) the eigenvalues of H, sorted in decreasing order: $h_d = L \geq h_{d-1} \geq \cdots \geq h_2 \geq h_1 = \mu > 0$, then Eq. (3) may be rewritten as:

$$P^\top \eta_{n+1} = (I - \alpha \mathrm{Diag}\,(h))\,P^\top \eta_n + (I - \beta \mathrm{Diag}\,(h))\,\left(P^\top \eta_n - P^\top \eta_{n-1}\right). \tag{5}$$

Thus there is no interaction between the different eigenspaces and we may consider, for the analysis only, d different recursions with $\eta_n^i = p_i^\top \eta_n$, $i \in \{1, ..., d\}$, where $p_i \in \mathbb{R}^d$ is the i-th column of P:

$$\eta_{n+1}^i = (1 - \alpha h_i)\,\eta_n^i + (1 - \beta h_i)\,\left(\eta_n^i - \eta_{n-1}^i\right). \tag{6}$$

3.1 Characteristic polynomial and eigenvalues

In this section, we consider a fixed $i \in \{1, \ldots, d\}$ and study the stability in the corresponding eigenspace. This linear dynamical system may be analyzed by studying the eigenvalues of the 2×2-matrix $F_i = \begin{pmatrix} 2 - (\alpha+\beta)h_i & \beta h_i - 1 \\ 1 & 0 \end{pmatrix}$. These eigenvalues are the roots of its characteristic polynomial which is:

$$\det(XI - F_i) = \det\left(X\left(X - 2 + (\alpha+\beta)h_i\right) + 1 - \beta h_i\right) = X^2 - 2X\Big(1 - \Big(\frac{\alpha+\beta}{2}\Big)h_i\Big) + 1 - \beta h_i.$$

To compute the roots of the second-order polynomial, we compute its reduced discriminant:

$$\Delta_i = \Big(1 - \Big(\frac{\alpha+\beta}{2}\Big)h_i\Big)^2 - 1 + \beta h_i = h_i\Big(\Big(\frac{\alpha+\beta}{2}\Big)^2 h_i - \alpha\Big).$$

Depending on the sign of the discriminant Δ_i, there will be two real distinct eigenvalues ($\Delta_i > 0$), two complex conjugate eigenvalues ($\Delta_i < 0$) or a single real eigenvalue ($\Delta_i = 0$).

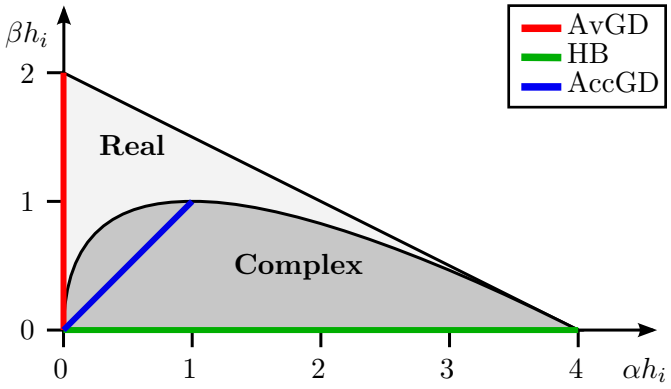

Figure 1: Area of stability of the algorithm, with the three traditional algorithms represented. In the interior of the triangle, the convergence is linear.

We will now study the sign of Δ_i. In each different case, we will determine under what conditions on α and β the modulus of the eigenvalues is less than one, which means that the iterates $(\eta_n^i)_n$ remain bounded and the iterates $(\theta_n)_n$ converge to θ_*. We may then compute function values as $f(\theta_n) - f(\theta_*) = \frac{1}{2n^2}\sum_{i=1}^d (\eta_n^i)^2 h_i = \frac{1}{2}\sum_{i=1}^d (\phi_n^i)^2 h_i$.

The various regimes are summarized in Figure 1: there is a triangle of values of $(\alpha h_i, \beta h_i)$ for which the algorithm remains stable (i.e., the iterates $(\eta_n)_n$ do not diverge), with either complex or real eigenvalues. In the following lemmas (see proof in Appendix C), we provide a detailed analysis that leads to Figure 1.

Lemma 1 (Real eigenvalues)**.** *The discriminant Δ_i is strictly positive and the algorithm is stable if and only if*

$$\alpha \geq 0, \qquad \alpha + 2\beta \leq 4/h_i, \qquad \alpha + \beta > 2\sqrt{\alpha/h_i}.$$

We then have two real roots $r_i^{\pm} = r_i \pm \sqrt{\Delta_i}$, with $r_i = 1 - (\frac{\alpha+\beta}{2})h_i$. Moreover, we have:

$$(\phi_n^i)^2 h_i = \frac{(\phi_1^i)^2 h_i}{4n^2} \frac{\left[(r_i + \sqrt{\Delta_i})^n - (r_i - \sqrt{\Delta_i})^n\right]^2}{\Delta_i}. \tag{7}$$

Therefore, for real eigenvalues, $((\phi_n^i)^2 h_i)_n$ will converge to 0 at a speed of $O(1/n^2)$ however the constant Δ_i may be arbitrarily small (and thus the scaling factor arbitrarily large). Furthermore we have linear convergence if the inequalities in the lemmas are strict.

Lemma 2 (Complex eigenvalues)**.** *The discriminant Δ_i is stricly negative and the algorithm is stable if and only if*

$$\alpha \geq 0, \qquad \beta \geq 0, \qquad \alpha + \beta < \sqrt{\alpha/h_i}.$$

We then have two complex conjugate eigenvalues: $r_i^{\pm} = r_i \pm \sqrt{-1}\sqrt{-\Delta_i}$. Moreover, we have:

$$(\phi_n^i)^2 h_i = \frac{(\phi_1^i)^2}{n^2} \frac{\sin^2(\omega_i n)}{\left(\alpha - (\frac{\alpha+\beta}{2})^2 h_i\right)} \rho^{2n}. \tag{8}$$

with $\rho_i = \sqrt{1 - \beta h_i}$, and ω_i defined through $\sin(\omega_i) = \sqrt{-\Delta_i}/\rho_i$ and $\cos(\omega_i) = r_i/\rho_i$.

Therefore, for complex eigenvalues, there is a linear convergence if the inequalities in the lemma are strict. Moreover, $((\phi_n^i)^2 h_i)_n$ oscillates to 0 at a speed of $O(1/n^2)$ even if h_i is arbitrarily small.

Coalescing eigenvalues. When the discriminant goes to zero in the explicit formulas of the real and complex cases, both the denominator and numerator of $((\phi_n^i)^2 h_i)_n$ will go to zero. In the limit case, when the discriminant is equal to zero, we will have a double real eigenvalue. This happens for $\beta = 2\sqrt{\alpha/h_i} - \alpha$. Then the eigenvalue is $r_i = 1 - \sqrt{\alpha h_i}$, and the algorithm is stable for $0 < \alpha < 4/h_i$, we then have $(\phi_n^i)^2 h_i = h_i (\phi_1^i)^2 (1 - \sqrt{\alpha h_i})^{2(n-1)}$. This can be obtained by letting Δ_i goes to 0 in the real and complex cases (see also Appendix C.3).

Summary. To conclude the iterate $(\eta_n^i)_n = (n(\theta_n^i - \theta_*^i))_n$ will be stable for $\alpha \in [0, 4/h_i]$ and $\beta \in [0, 2/h_i - \alpha/2]$. According to the values of α and β this iterate will have a different behavior. In the complex case, the roots are complex conjugate with magnitude $\sqrt{1 - \beta h_i}$. Thus, when $\beta > 0$, $(\eta_n^i)_n$ will converge to 0, oscillating, at rate $\sqrt{1 - \beta h_i}$. In the real case, the two roots are real and distinct. However the product of the two roots is equal to $\sqrt{1 - \beta h_i}$, thus one will have a higher magnitude and $(\eta_n^i)_n$ will converges to 0 at rate higher than in the complex case (as long as α and β belong to the interior of the stability region).

Finally, for a given quadratic function f, all the d iterates $(\eta_n^i)_n$ should be bounded, therefore we must have $\alpha \in [0, 4/L]$ and $\beta \in [0, 2/L - \alpha/2]$. Then, depending on the value of h_i, some eigenvalues may be complex or real.

3.2 Classical examples

For particular choices of α and β, displayed in Figure 1, the eigenvalues are either all real or all complex, as shown in the table below.

	Av-GD	Acc-GD	Heavy ball
α	0	γ	γ
β	γ	γ	0
Δ_i	$(\gamma h_i)^2$	$-\gamma h_i (1 - \gamma h_i)$	$-\gamma h_i (1 - \frac{\gamma h_i}{4})$
$r_i^{\pm}$	$1,\ 1 - \gamma h_i$	$\sqrt{1 - \gamma h_i} e^{\pm i \omega_i}$	$e^{\pm i \omega_i}$
$\cos(\omega_i)$		$\sqrt{1 - \gamma h_i}$	$1 - \frac{\gamma}{2} h_i$
ρ_i		$\sqrt{1 - \gamma h_i}$	1

Averaged gradient descent loses linear convergence for strongly-convex problems, because $r_i^+ = 1$ for all eigensubspaces. Similarly, the heavy ball method is not adaptive to strong convexity because $\rho_i = 1$. However, accelerated gradient descent, although designed for non-strongly-convex problems, is adaptive because $\rho_i = \sqrt{1 - \gamma h_i}$ depends on h_i while α and β do not. These last two algorithms have an oscillatory behavior which can be observed in practice and has been already studied (Su et al., 2014).

Note that all the classical methods choose step-sizes α and β either having all the eigenvalues real either complex; whereas we will see in Section 4, that it is significant to combine both behaviors in presence of noise.

3.3 General bound

Even if the exact formulas in Lemmas 1 and 2 are computable, they are not easily interpretable. In particular when the two roots become close, the denominator will go to zero, which prevents from bounding them easily. When we further restrict the domain of (α, β), we can always bound the iterate by the general bound (see proof in Appendix D):

Theorem 2. *For $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$, we have*

$$(\eta_n^i)^2 \leq \min\left\{ \frac{2(\eta_1^i)^2}{\alpha h_i}, \frac{8(\eta_1^i)^2 n}{(\alpha+\beta)h_i}, \frac{16(\eta_1^i)^2}{(\alpha+\beta)^2 h_i^2} \right\}. \tag{9}$$

These bounds are shown by dividing the set of (α, β) in three regions where we obtain specific bounds. They do not depend on the regime of the eigenvalues (complex or real); this enables us to get the following general bound on the function values, our main result for the deterministic case.

Corollary 1. *For $\alpha \leq 1/L$ and $0 \leq \beta \leq 2/L - \alpha$:*

$$f(\theta_n) - f(\theta_*) \leq \min\left\{ \frac{\|\theta_0 - \theta_*\|^2}{\alpha n^2}, \frac{4\|\theta_0 - \theta_*\|^2}{(\alpha+\beta)n} \right\}. \tag{10}$$

We can make the following observations:

- The first bound $\frac{\|\theta_0-\theta_*\|^2}{\alpha n^2}$ corresponds to the traditional acceleration result, and is only relevant for $\alpha > 0$ (that is, for Nesterov acceleration and the heavy-ball method, but not for averaging). We recover the traditional convergence rate of second-order methods for quadratic functions in the singular case, such as conjugate gradient (Polyak, 1987, Section 6.1).
- While the result above focuses on function values, like most results in the non-strongly convex case, the distance to optimum $\|\theta_n - \theta_*\|^2$ typically does not go to zero (although it remains bounded in our situation).
- When $\alpha = 0$ (averaged gradient descent), then the second bound $\frac{4\|\theta_0-\theta_*\|^2}{(\alpha+\beta)n}$ provides a convergence rate of $O(1/n)$ if no assumption is made regarding the starting point θ_0, while the last bound of Theorem 2 would lead to a bound $\frac{8\|H^{-1/2}(\theta_0-\theta_*)\|^2}{(\alpha+\beta)^2 n^2}$, that is a rate of $O(1/n^2)$, only for some starting points.
- As shown in Appendix E by exhibiting explicit sequences of quadratic functions, the inverse dependence in αn^2 and $(\alpha+\beta)n$ in Eq. (10) is not improvable.

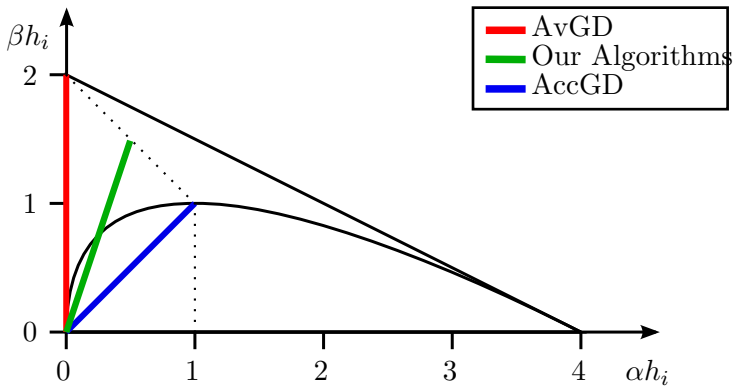

Figure 2: Trade-off between averaged and accelerated methods for noisy gradients.

4 Quadratic Optimization with Additive Noise

In many practical situations, the gradient of f is not available for the recursion in Eq. (4), but only a noisy version. In this paper, we only consider additive uncorrelated noise with finite variance.

4.1 Stochastic difference equation

We now assume that the true gradient is not available and we rather have access to a noisy oracle for the gradient of f. In Eq. (4), we assume that the oracle outputs a noisy gradient $f'\big(\frac{n(\alpha+\beta)}{n\alpha+\beta}\theta_n - \frac{(n-1)\beta}{n\alpha+\beta}\theta_{n-1}\big) - \varepsilon_{n+1}$. The noise (ε_n) is assumed to be uncorrelated zero-mean with bounded covariance, i.e., $\mathbb{E}[\varepsilon_n \otimes \varepsilon_m] = 0$ for all $n \neq m$ and $\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] \preccurlyeq C$, where $A \preccurlyeq B$ means that $B - A$ is positive semi-definite.

For quadratic functions, for the reduced variable $\eta_n = n\phi_n = n(\theta_n - \theta_*)$, we get:

$$\eta_{n+1} = (I - \alpha H)\eta_n + (I - \beta H)(\eta_n - \eta_{n-1}) + [n\alpha + \beta]\varepsilon_{n+1}. \tag{11}$$

Note that algorithms with $\alpha \neq 0$ will have an important level of noise because of the term $n\alpha\varepsilon_{n+1}$. We denote by $\xi_{n+1} = \begin{pmatrix} [n\alpha+\beta]\varepsilon_{n+1} \\ 0 \end{pmatrix}$ and we now have the recursion:

$$\Theta_{n+1} = F\Theta_n + \xi_{n+1}, \tag{12}$$

which is a standard noisy linear dynamical system (see, e.g., Arnold, 1998) with uncorrelated noise process (ξ_n). We may thus express Θ_n directly as $\Theta_n = F^n\Theta_0 + \sum_{k=1}^n F^{n-k}\xi_k$, and its expected second-order moment as, $\mathbb{E}\big(\Theta_n\Theta_n\big)^\top = F^n\Theta_0\Theta_0^\top(F^n)^\top + \sum_{k=1}^n F^{n-k}\mathbb{E}\big(\xi_k\xi_k^\top\big)(F^{n-k})^\top$. In order to obtain the expected excess cost function, we simply need to compute $\operatorname{tr}\begin{pmatrix} 0 & H \\ 0 & 0 \end{pmatrix}\mathbb{E}\big(\Theta_n\Theta_n\big)^\top$, which thus decomposes as a term that only depends on initial conditions (which is exactly the one computed and studied in Section 3.3), and a new term that depends on the noise.

4.2 Convergence result

For a quadratic function f with arbitrarily small eigenvalues and uncorrelated noise with finite covariance, we obtain the following convergence result (see proof in Appendix F); since we will allow the parameters α and β to depend on the time we stop the algorithm, we introduce the horizon N:

Theorem 3 (Convergence rates with noisy gradients)**.** *With $\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] = C$ for all $n \in \mathbb{N}$, for $\alpha \leq \frac{1}{L}$ and $0 \leq \beta \leq \frac{2}{L} - \alpha$. Then for any $N \in \mathbb{N}$, we have:*

$$\mathbb{E}f(\theta_N) - f(\theta_*) \leq \min\left\{ \frac{\|\theta_0 - \theta_*\|^2}{\alpha N^2} + \frac{(\alpha N + \beta)^2}{\alpha N} \operatorname{tr}(C), \frac{4\|\theta_0 - \theta_*\|^2}{(\alpha+\beta)N} + \frac{4(\alpha N + \beta)^2}{\alpha + \beta} \operatorname{tr}(C) \right\}. \quad (13)$$

We can make the following observations:

- Although we only provide an upper-bound, the proof technique relies on direct moment computations in each eigensubspace with few inequalities, and we conjecture that the scalings with respect to n are tight.
- For $\alpha = 0$ and $\beta = 1/L$ (which corresponds to averaged gradient descent), the second bound leads to $\frac{4L\|\theta_0 - \theta_*\|^2}{N} + \frac{4\operatorname{tr}(C)}{L}$, which is bounded but not converging to zero. We recover a result from Bach and Moulines (2011, Theorem 1).
- For $\alpha = \beta = 1/L$ (which corresponds to Nesterov's acceleration), the first bound leads to $\frac{L\|\theta_0 - \theta_*\|^2}{N^2} + \frac{(N+1)\operatorname{tr}(C)}{L}$, and our bound suggests that the algorithm diverges, which we have observed in our experiments in Appendix A.
- For $\alpha = 0$ and $\beta = 1/L\sqrt{N}$, the second bound leads to $\frac{4L\|\theta_0 - \theta_*\|^2}{\sqrt{N}} + \frac{4\operatorname{tr}(C)}{L\sqrt{N}}$, and we recover the traditional rate of $1/\sqrt{N}$ for stochastic gradient in the non-strongly-convex case.
- When the values of the bias and the variance are known we can choose α and β such that the trade-off between the bias and the variance is optimal in our bound, as the following corrollary shows. Note that in the bound below, taking a non zero β enables the bias term to be adaptive to hidden strong-convexity.

Corollary 2. *For $\alpha = \min\left\{ \frac{\|\theta_0 - \theta_*\|}{2\sqrt{\operatorname{tr} C} N^{3/2}}, 1/L \right\}$ and $\beta \in [0, \min\{N\alpha, 1/L\}]$, we have:*

$$\mathbb{E}f(\theta_N) - f(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{N^2} + \frac{4\sqrt{\operatorname{tr} C}\|\theta_0 - \theta_*\|}{\sqrt{N}}.$$

4.3 Structured noise and least-square regression

When only the noise total variance $\operatorname{tr}(C)$ is considered, as shown in Section 4.4, Corollary 2 recover existing (more general) results. Our framework however leads to improved result for *structured noise processes* frequent in machine learning, in particular in least-squares regression which we now consider but this goes beyond (see, e.g. Bach and Moulines, 2013).

Assume we observe independent and identically distributed pairs $(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ and we want to minimize the expected loss $f(\theta) = \frac{1}{2}\mathbb{E}[(y_n - \langle \theta, x_n \rangle)^2]$. We denote by $H = \mathbb{E}(x_n \otimes x_n)$ the covariance matrix which is assumed invertible. The global minimum of f is attained at $\theta_* \in \mathbb{R}^d$ defined as before and we denote by $r_n = y_n - \langle \theta_*, x_n \rangle$ the statistical noise, which we assume bounded by σ. We have $\mathbb{E}[r_n x_n] = 0$. In an online setting, we observe the gradient $(x_n \otimes x_n)(\theta - \theta_*) - r_n x_n$, whose expectation is the gradient $f'(\theta)$. This corresponds to a noise in the gradient of $\varepsilon_n = (H - x_n \otimes x_n)(\theta - \theta_*) + r_n x_n$. Given θ, if the data (x_n, y_n) are almost surely bounded, the covariance matrix of this noise is bounded by a constant times H. This suggests to characterize the noise convergence by $\operatorname{tr}(CH^{-1})$, which is bounded even though H has arbitrarily small eigenvalues.

However, our result will not apply to stochastic gradient descent (SGD) for least-squares, because of the term $(H - x_n \otimes x_n)(\theta - \theta_*)$ which depends on θ, but to a "semi-stochastic" recursion where the noisy gradient is $H(\theta - \theta_*) - r_n x_n$, with a noise process $\varepsilon_n = r_n x_n$, which is such that $\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] \preccurlyeq \sigma^2 H$, and has been used by Bach and Moulines (2011) and Dieuleveut and Bach (2014) to prove results on regular stochastic gradient descent. We conjecture that our algorithm (and results) also applies in the regular SGD case, and we provide encouraging experiments in Section 5.

For this particular structured noise we can take advantage of a large β:

Theorem 4 (Convergence rates with structured noisy gradients)**.** *Let $\alpha \leq \frac{1}{L}$ and $0 \leq \beta \leq \frac{3}{2L} - \frac{\alpha}{2}$. For any $N \in \mathbb{N}$, $\mathbb{E}f(\theta_N) - f(\theta_*)$ is upper-bounded by:*

$$\min\left\{\frac{\|\theta_0 - \theta_*\|^2}{N^2\alpha} + \frac{(\alpha N + \beta)^2}{\alpha\beta N^2}\operatorname{tr}(CH^{-1}), \frac{4L\|\theta_0 - \theta_*\|^2}{(\alpha+\beta)N} + \frac{8(\alpha N + \beta)^2 \operatorname{tr}(CH^{-1})}{(\alpha+\beta)^2 N}\right\}. \quad (14)$$

We can make the following observations:

- For $\alpha = 0$ and $\beta = 1/L$ (which corresponds to averaged gradient descent), the second bound leads to $\frac{4L\|\theta_0 - \theta_*\|^2}{N} + \frac{8\operatorname{tr}(CH^{-1})}{N}$. We recover a result from Bach and Moulines (2013, Theorem 1). Note that when $C \preccurlyeq \sigma^2 H$, $\operatorname{tr}(CH^{-1}) \leqslant \sigma^2 d$.
- For $\alpha = \beta = 1/L$ (which corresponds to Nesterov's acceleration), the first bound leads to $\frac{L\|\theta_0 - \theta_*\|^2}{N^2} + \operatorname{tr}(CH^{-1})$, which is bounded but not converging to zero (as opposed to the the unstructured noise where the algorithm may diverge).
- For $\alpha = 1/(LN^a)$ with $0 \leq a \leq 1$ and $\beta = 1/L$, the first bound leads to $\frac{L\|\theta_0 - \theta_*\|^2}{N^{2-a}} + \frac{\operatorname{tr}(CH^{-1})}{N^a}$. We thus obtain an explicit bias-variance trade-off by changing the value of a.
- When the values of the bias and the variance are known we can choose α and β with an optimized trade-off, as the following corrollary shows:

Corollary 3. *For $\alpha = \min\left\{\frac{\|\theta_0 - \theta_*\|}{\sqrt{L\operatorname{tr}(CH^{-1})N}}, 1/L\right\}$ and $\beta = \min\{N\alpha, 1/L\}$ we have:*

$$\mathbb{E}f(\theta_N) - f(\theta_*) \leq \max\left\{\frac{5\operatorname{tr}(CH^{-1})}{N}, \frac{5\sqrt{\operatorname{tr}(CH^{-1})L}\|\theta_0 - \theta_*\|}{N}, \frac{2\|\theta_0 - \theta_*\|^2 L}{N^2}\right\}. \quad (15)$$

4.4 Related work

Acceleration and noisy gradients. Several authors (Lan, 2012; Hu et al., 2009; Xiao, 2010) have shown that using a step-size proportional to $1/N^{3/2}$ accelerated methods with noisy gradients lead to the same convergence rate of $O\big(\frac{L\|\theta_0-\theta_*\|^2}{N^2} + \frac{\|\theta_0-\theta_*\|\sqrt{\operatorname{tr}(C)}}{\sqrt{N}}\big)$ than in Corollary 2, for smooth functions. Thus, for unstructured noise, our analysis provides insights in the behavior of second-order algorithms, without improving bounds. We get significant improvements for structured noises.

Least-squares regression. When the noise is structured as in least-square regression and more generally in linear supervised learning, Bach and Moulines (2011) have shown that using averaged stochastic gradient descent with constant step-size leads to the convergence rate of $O\big(\frac{L\|\theta_0-\theta_0\|^2}{N}+\frac{\sigma^2 d}{N}\big)$. It has been highlighted by Défossez and Bach (2014) that the bias term $\frac{L\|\theta_0-\theta_*\|^2}{N}$ may often be the dominant one in practice. Our result in Corollary 3 leads to an improved bias term in $O(1/N^2)$ with the price of a potentially slightly worse constant in the variance term. However, with optimal constants in Corollary 3, the new algorithm is always an improvement over averaged stochastic gradient descent in all situations. If constants are unknown, we may use $\alpha = 1/(LN^a)$ with $0 \leq a \leq 1$ and $\beta = 1/L$ and we choose a depending on the emphasis we want to put on bias or variance.

Minimax convergence rates. For noisy quadratic problems, the convergence rate nicely decomposes into two terms, a bias term which corresponds to the noiseless problem and the variance term which corresponds to a problem started at θ_*. For each of these two terms, lower bounds are known. For the bias term, if $N \leq d$, then the lower bound is, up to constants, $L\|\theta_0 - \theta_*\|^2/N^2$ (Nesterov, 2004, Theorem 2.1.7). For the variance term, for the general noisy gradient situation, we show in Appendix H that for $N \leq d$, it is $(\operatorname{tr} C)/(L\sqrt{N})$, while for least-squares regression, it is $\sigma^2 d/N$ (Tsybakov, 2003). Thus, for the two situations, we attain the two lower bounds *simultaneously* for situations where respectively $L\|\theta_0 - \theta_*\|^2 \leq (\operatorname{tr} C)/L$ and $L\|\theta_0 - \theta_*\|^2 \leq d\sigma^2$. It remains an open problem to achieve the two minimax terms in all situations.

Other algorithms as special cases. We also note as shown in Appendix G that in the special case of quadratic functions, the algorithms of Lan (2012); Hu et al. (2009); Xiao (2010) could be unified into our framework (although they have significantly different formulations and justifications in the smooth case).

5 Experiments

In this section, we illustrate our theoretical results on synthetic examples. We consider a matrix H that has random eigenvectors and eigenvalues $1/k^m$, for $k = 1, \dots, d$ and $m \in \mathbb{N}$. We take a random optimum θ_* and a random starting point θ_0 such that $r = \|\theta_0 - \theta_*\| = 1$ (unless otherwise specified). In Appendix A, we illustrate the noiseless results

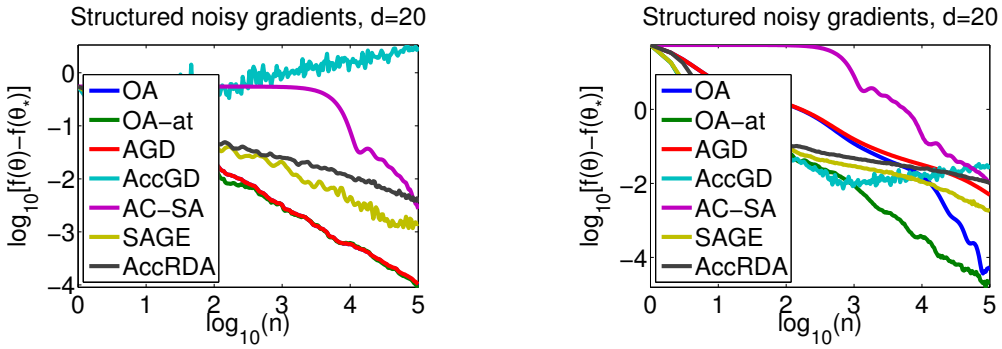

Figure 3: Quadratic optimization with regression noise. Left $\sigma = 1$, $r = 1$. Right $\sigma = 0.1$, $r = 10$.

of Section 3, in particular the oscillatory behaviors and the influence of all eigenvalues, as well as unstructured noisy gradients. In this section, we focus on noisy gradients with structured noise (as described in Section 4.3), where our new algorithms show significant improvements.

We compare our algorithm to other stochastic accelerated algorithms, that is, AC-SA (Lan, 2012), SAGE (Hu et al., 2009) and Acc-RDA (Xiao, 2010) which are presented in Appendix G. For all these algorithms (and ours) we take the optimal step-sizes defined in these papers. We show results averaged over 10 replications.

Homoscedastic noise. We first consider an i.i.d. zero mean noise whose covariance matrix is proportional to H. We also consider a variant of our algorithm with an any-time step-size function of n rather than N (for which we currently have no proof of convergence). In Figure 3, we take into account two different set-ups. In the left plot, the variance dominates the bias (with $r = \|\theta_0 - \theta_*\| = \sigma$). We see that (a) Acc-GD does not converge to the optimum but does not diverge either, (b) Av-GD and our algorithms achieve the optimal rate of convergence of $O(\sigma^2 d/n)$, whereas (c) other accelerated algorithms only converge at rate $O(1/\sqrt{n})$. In the right plot, the bias dominates the variance ($r = 10$ and $\sigma = 0.1$). In this situation our algorithm outperforms all others.

Application to least-squares regression. We now see how these algorithms behave for least-squares regressions and the regular (non-homoscedastic) stochastic gradients described in Section 4.3. We consider normally distributed inputs. The covariance matrix H is the same as before. The outputs are generated from a linear function with homoscedatic noise with a signal-to-noise ratio of σ. We consider $d = 20$. We show results averaged over 10 replications. In Figure 4, we consider again a situation where the bias dominates (left) and vice versa (right). We see that our algorithm has the same good behavior than in the homoscedastic noise case and we conjecture that our bounds also hold in this situation.

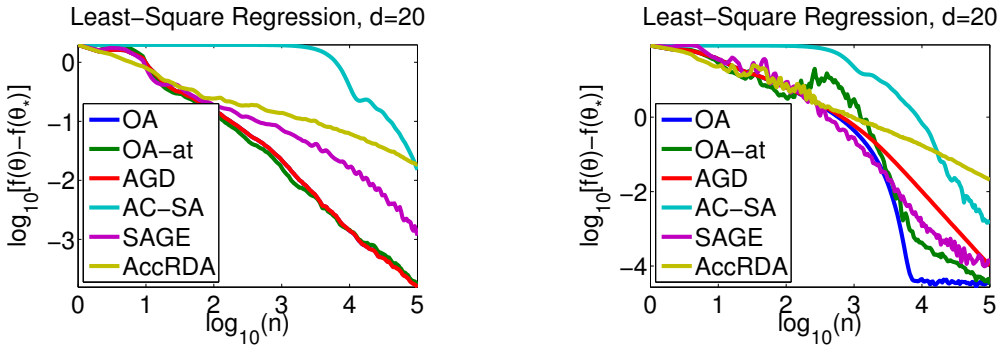

Figure 4: Least-Square Regression. Left $\sigma = 1$, $r = 1$. Right $\sigma = 0.1$, $r = 10$.

6 Conclusion

We have provided a joint analysis of averaging and acceleration for non-strongly-convex quadratic functions in a single framework, both with noiseless and noisy gradients. This allows to define a class of algorithms that can benefit simultaneously of the known improvements of averaging and accelerations: faster forgetting of initial conditions (for acceleration), and better robustness to noise when the noise covariance is proportional to the Hessian (for averaging).

Our current analysis of our class of algorithms in Eq. (4), that considers two different affine combinations of previous iterates (instead of one for traditional acceleration), is limited to quadratic functions; an extension of its analysis to all smooth or self-concordant-like functions would widen its applicability. Similarly, an extension to least-squares regression with natural heteroscedastic stochastic gradient, as suggested by our simulations, would be an interesting development.

Acknowledgements

This work was partially supported by the MSR-Inria Joint Centre and a grant by the European Research Council (SIERRA project 239993). The authors would like to thank Aymeric Dieuleveut for helpful discussions.

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *Information Theory, IEEE Transactions on*, 58(5):3235–3249, 2012.

L. Arnold. *Random dynamical systems*. Springer Monographs in Mathematics. Springer-Verlag, 1998.

F. Bach and E. Moulines. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Advances in Neural Information Processing Systems*, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, December 2013.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3): 1171–1183, 2008.

A. Défossez and F. Bach. Constant step size least-mean-square: Bias-variance trade-offs and optimal sampling distributions. Technical Report 1412.0156, arXiv, 2014.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2, Ser. A):37–75, 2014.

A. Dieuleveut and F. Bach. Non-parametric Stochastic Approximation with Large Step sizes. Technical Report 1408.0361, arXiv, August 2014.

C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, 2009.

G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133 (1-2, Ser. A):365–397, 2012.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140 (1, Ser. B):125–161, 2013.

B. O'Donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, pages 1–18, 2013.

J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*, volume 30 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. {*USSR*} *Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

B. T. Polyak. *Introduction to Optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

M. Schmidt, N. Le Roux, and F. Bach. Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization. In *Advances in Neural Information Processing Systems*, December 2011.

W. Su, S. Boyd, and E. Candes. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. In *Advances in Neural Information Processing Systems*, 2014.

A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2003.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.

A Additional experimental results

In this appendix, we provide additional experimental results to illustrate our theoretical results.

A.1 Deterministic convergence

Comparaison for $d = 1$. In Figure 5, we minimize a one-dimensional quadratic function $f(\theta) = \frac{1}{2}\theta^2$ for a fixed step-size $\alpha = 1/10$ and different step-sizes β. In the left plot, we compare Acc-GD, HB and Av-GD. We see that HB and Acc-GD both oscillate and that Acc-GD leverages strong convexity to converge faster. In the right plot, we compare the behavior of the algorithm for different values of β. We see that the optimal rate is achieved for $\beta = \beta_*$ defined to be the one for which there is a double coalescent eigenvalue, where the convergence is linear at speed $O(1 - \sqrt{\alpha L})^n$. When $\beta > \beta_*$, we are in the real case and when $\beta < \beta_*$ the algorithm oscillates to the solution.

Comparison between the different eigenspaces. Figure 6 shows interactions between different eigenspaces. In the left plot, we optimize a quadratic function of dimension $d = 2$. The first eigenvalue is $L = 1$ and the second is $\mu = 2^{-8}$. For Av-GD the convergence is of order $O(1/n)$ since the problem is "not" strongly convex (i.e., not appearing as strongly convex since $n\mu$ remains small). The convergence is at the beginning the same for HB and Acc-GD, with oscillation at speed $O(1/n^2)$, since the small eigenvalue prevents Acc-GD from having a linear convergence. Then for large n, the convergence becomes linear for Acc-GD, since μn becomes large. In the right plot, we optimize a quadratic function in dimension $d = 5$ with eigenvalues from 1 to 0.1. We show the function values of the projections of the iterates η_n on the different eigenspaces. We see that high eigenvalues first dominate, but converge quickly to zero, whereas small ones keep oscillating, and converge more slowly.

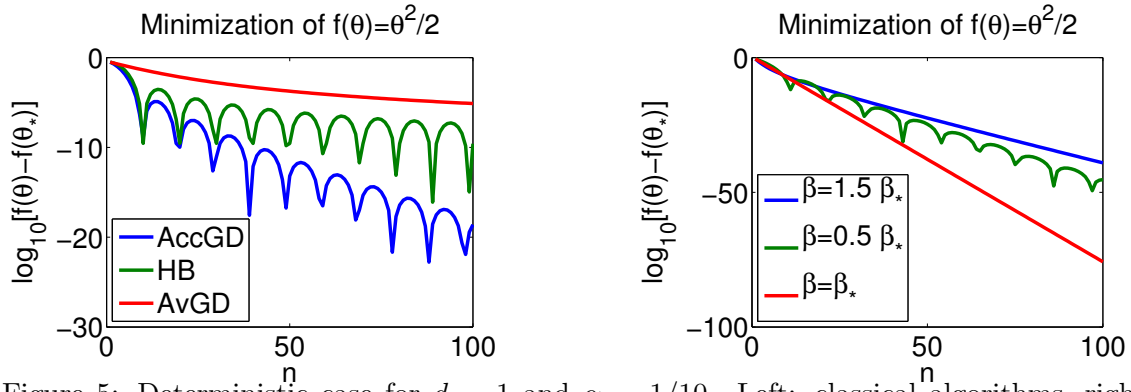

Figure 5: Deterministic case for $d = 1$ and $\alpha = 1/10$. Left: classical algorithms, right: different oscillatory behaviors.

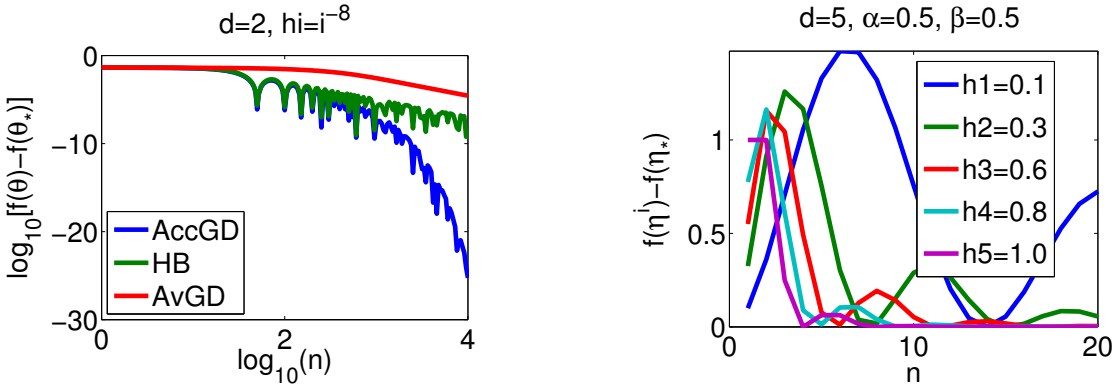

Figure 6: Left: Deterministic quadratic optimization for $d = 2$. Right: Function value of the projection of the iterate on the different eigenspaces ($d = 5$).

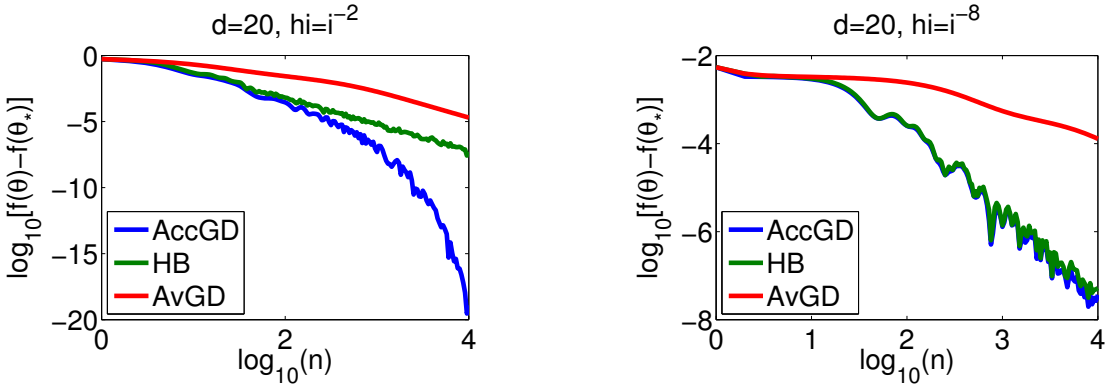

Figure 7: Deterministic case for $d = 20$ and $\gamma = 1/10$. Left: $m = 2$. Right: $m = 8$.

Comparison for $d = 20$. In Figure 7, we optimize two 20-dimensional quadratic functions with different eigenvalues with Av-GD, HB and Acc-GD for a fixed step-size $\gamma = 1/10$. In the left plot, the eigenvalues are $1/k^2$ and in the right one, they are $1/k^8$, for $k = 1, \dots, d$. We see that in both cases, Av-GD converges at a rate of $O(1/n)$ and HB at a rate of

$O(1/n^2)$. For Acc-GD the convergence is linear when μ is large (left plot) and becomes sublinear at a rate of $O(1/n^2)$ when μ becomes small (right plot).

A.2 Noisy convergence with unstructured additive noise

We optimize the same quadratic function, but now with noisy gradients. We compare our algorithm to other stochastic accelerated algorithms, that is, AC-SA (Lan, 2012), SAGE (Hu et al., 2009) and Acc-RDA (Xiao, 2010), which are presented in Appendix G. For all these algorithms (and ours) we take the optimal step-sizes defined in these papers. We plot the results averaged over 10 replications.

We consider in Figure 8 an i.i.d. zero mean noise of variance $C = I$. We see that all the accelerated algorithms achieve the same precision whereas Av-GD with constant step-size does not converge and Acc-Gd diverges. However SAGE and AC-SA are anytime algorithms and are faster at the beginning since their step-sizes are decreasing and not a constant (with respect to n) function of the horizon N.

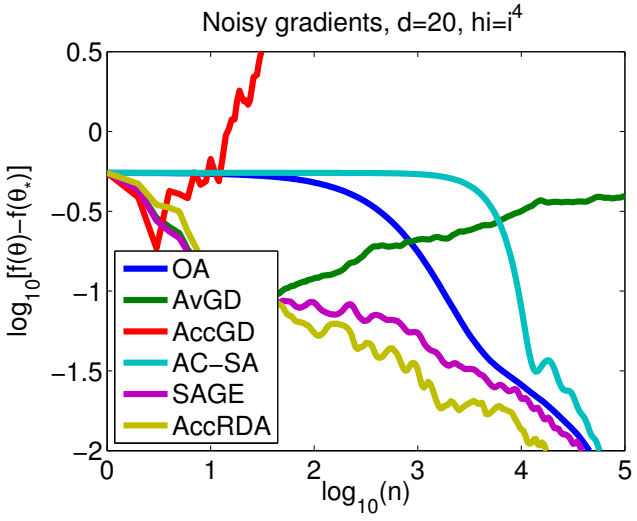

Figure 8: Quadratic optimization with additive noise.

B Proofs of Section 2

B.1 Proof of Theorem 1

Let $(P_n, Q_n, R_n) \in (\mathbb{R}[X])^3$ for all $n \in \mathbb{N}$ be a sequence of polynomials. We consider the iterates defined for all $n \in \mathbb{N}^*$ by

$$\theta_{n+1} = P_n(H)\theta_n + Q_n(H)\theta_{n-1} + R(H)q,$$

started from $\theta_0 = \theta_1 \in \mathbb{R}^d$. The θ_*-stationarity property gives for $n \in \mathbb{N}^*$:

$$\theta_* = P_n(H)\theta_* + Q_n(H)\theta_* + R_n(H)q.$$

Since $\theta_* = H^{-1}q$ we get for all $q \in \mathbb{R}^d$

$$H^{-1}q = P_n(H)H^{-1}q + Q_n(H)H^{-1}q + R_n(H)q.$$

For all $\tilde{q} \in \mathbb{R}^d$ we apply this relation to vectors $q = H\tilde{q}$:

$$\tilde{q} = P_n(H)\tilde{q} + Q_n(H)\tilde{q} + R_n(H)H\tilde{q} \quad \forall \tilde{q} \in \mathbb{R}^d,$$

and we get

$$I = P_n(H) + Q_n(H) + R_n(H)H \quad \forall n \in \mathbb{N}^*.$$

Therefore there are polynomials $(\bar{P}_n, \bar{Q}_n) \in (\mathbb{R}[X])^2$ and $q_n \in \mathbb{R}$ for all $n \in \mathbb{N}^*$ such that we have for all $n \in \mathbb{N}$:

$$\begin{aligned} P_n(X) &= (1-q_n)I + X\bar{P}_n(X) \\ Q_n(X) &= q_n I + X\bar{Q}_n(X) \\ R_n(X) &= -(\bar{P}_n(X) + \bar{Q}_n(X)). \end{aligned} \tag{16}$$

The n-scalability property means that there are polynomials $(P, Q) \in (\mathbb{R}[X])^2$ independent of n such that:

$$\begin{aligned} P_n(X) &= \frac{n}{n+1}P(X), \\ Q_n(X) &= \frac{n-1}{n+1}Q(X). \end{aligned}$$

And in connection with Eq. (16) we can rewrite P and Q as:

$$\begin{aligned} P(X) &= \bar{p} + X\bar{P}(X), \\ Q(X) &= \bar{q} + X\bar{Q}(X), \end{aligned}$$

with $(\bar{p}, \bar{q}) \in \mathbb{R}^2$ and $(\bar{P}, \bar{Q}) \in (\mathbb{R}[X])^2$. Thus for all $n \in \mathbb{N}$:

$$q_n = \frac{n-1}{n+1}\bar{q} \tag{17}$$

$$\begin{aligned} \bar{Q}_n(X) &= \frac{n-1}{n}Q(X) \\ \frac{n}{n+1}\bar{p} &= (1-q_n) \\ \bar{P}_n(X) &= \frac{n}{n+1}P(X). \end{aligned} \tag{18}$$

Eq. (17) and Eq. (18) give:

$$\frac{n}{n+1}\bar{p} = \left(1 - \frac{n-1}{n+1}\bar{q}\right).$$

Thus for $n = 1$, we have $\bar{p} = 2$. Then $-\frac{n-1}{n+1}\bar{q} = \frac{2n}{n+1} - 1 = \frac{n-1}{n+1}$ and $\bar{q} = -1$. Therefore

$$\begin{aligned} P_n(H) &= \frac{2n}{n+1}I + \frac{n}{n+1}\bar{P}(H)H \\ Q_n(H) &= -\frac{n-1}{n}I + \bar{Q}(H)H \\ R_n(H) &= -\left(\frac{n\bar{P}(H) + (n-1)\bar{Q}(H)}{n+1}\right). \end{aligned}$$

We let $\bar{A} = -(\bar{P} + \bar{Q})$ and $\bar{B} = \bar{Q}$ so that we have:

$$\begin{aligned} P_n(H) &= \frac{2n}{n+1}\left(I - \left(\frac{\bar{A}(H) + \bar{B}(H)}{2}\right)H\right) \\ Q_n(H) &= -\frac{n-1}{n}\left(I - \bar{B}(H)H\right) \\ R_n(H) &= \left(\frac{n\bar{A}(H) + \bar{B}(H)}{n+1}\right), \end{aligned}$$

and with $\phi_n = \theta_n - \theta_*$ for all $n \in \mathbb{N}$, the algorithm can be written under the form:

$$\phi_{n+1} = \left[I - \left(\frac{n}{n+1}\bar{A}(H) + \frac{1}{n+1}\bar{B}(H)\right)H\right]\phi_n + \left(1 - \frac{2}{n+1}\right)\left[I - \bar{B}(H)H\right](\phi_n - \phi_{n-1}).$$

B.2 Av-GD as two steps-algorithm

We show now that when the averaged iterate of Av-GD is seen as the main iterate we have that Av-GD with step-size $\gamma \in \mathbb{R}$ is equivalent to:

$$\theta_{n+1} = \frac{2n}{n+1}\theta_n - \frac{n-1}{n+1}\theta_{n-1} - \frac{\gamma}{n+1}f'\big(n\theta_n - (n-1)\theta_{n-1}\big).$$

We remind

$$\begin{aligned} \psi_{n+1} &= \psi_n - \gamma f'(\psi_n), \\ \theta_{n+1} &= \theta_n + \frac{1}{n+1}(\psi_{n+1} - \theta_n). \end{aligned}$$

Thus, we have:

$$\begin{aligned} \theta_{n+1} &= \theta_n + \frac{1}{n+1}(\psi_{n+1} - \theta_n) \\ &= \theta_n + \frac{1}{n+1}(\psi_n - \gamma f'(\psi_n) - \theta_n) \\ &= \theta_n + \frac{1}{n+1}(\theta_n + (n-1)(\theta_n - \theta_{n-1}) - \gamma f'(\theta_n + (n-1)(\theta_n - \theta_{n-1})) - \theta_n) \\ &= \frac{2n}{n+1}\theta_n - \frac{n-1}{n+1}\theta_{n-1} - \frac{\gamma}{n+1}f'(n\theta_n - (n-1)\theta_{n-1}). \end{aligned}$$

C Proof of Section 3

C.1 Proof of Lemma 1

The discriminant Δ_i is strictly positive when $\left(\frac{\alpha+\beta}{2}\right)^2 h_i - \alpha > 0$. This is always true for α strictly negative. For α positive and for $h_i \neq 0$, this is true for $|\frac{\alpha+\beta}{2}| > \sqrt{\alpha/h_i}$. Thus the discriminant Δ_i is strictly positive for

$$\begin{aligned} &\alpha < 0 \qquad \text{or} \\ &\alpha \geq 0 \qquad \text{and} \quad \left\{\beta < -\alpha - 2\sqrt{\alpha/h_i} \quad \text{or} \quad \beta > -\alpha + 2\sqrt{\alpha/h_i}\right\}. \end{aligned}$$

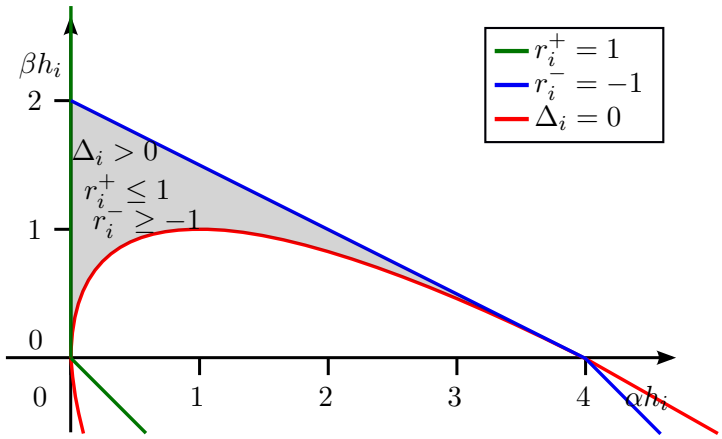

Figure 9: Stability in the real case, with all contraints plotted.

Then we determine when the modulus of the eigenvalues is less than one (which corresponds to $-1 \leq r_i^- \leq r_i^+ \leq 1$).

$$\begin{aligned}
r_i^+ \leq 1 \quad &\Leftrightarrow \quad \sqrt{h_i\left(\left(\frac{\alpha+\beta}{2}\right)^2 h_i - \alpha\right)} \leq \left(\frac{\beta+\alpha}{2}\right) h_i \\
&\Leftrightarrow \quad h_i\left(\left(\frac{\beta+\alpha}{2}\right)^2 h_i - \alpha\right) \leq \left[\left(\frac{\beta+\alpha}{2}\right) h_i\right]^2 \quad \text{and} \quad \frac{\alpha+\beta}{2} \geq 0 \\
&\Leftrightarrow \quad h_i\alpha \geq 0 \quad \text{and} \quad \frac{\alpha+\beta}{2} \geq 0 \\
&\Leftrightarrow \quad \alpha \geq 0 \quad \text{and} \quad \alpha+\beta \geq 0.
\end{aligned}$$

Moreover, we have :

$$\begin{aligned}
r_i^- \geq -1 \quad &\Leftrightarrow \quad \sqrt{h_i\left(\left(\frac{\beta+\alpha}{2}\right)^2 h_i - \alpha\right)} \leq 2 - \left(\frac{\beta+\alpha}{2}\right) h_i \\
&\Leftrightarrow \quad h_i\left(\left(\frac{\beta+\alpha}{2}\right)^2 h_i - \alpha\right) \leq \left[2 - \left(\frac{\beta+\alpha}{2}\right) h_i\right]^2 \text{ and } 2 - \left(\frac{\beta+\alpha}{2}\right) h_i \geq 0 \\
&\Leftrightarrow \quad h_i\left(\left(\frac{\beta+\alpha}{2}\right)^2 h_i - \alpha\right) \leq 4 - 4\left(\frac{\beta+\alpha}{2}\right) h_i + \left[\left(\frac{\beta+\alpha}{2}\right) h_i\right]^2 \text{ and } \left(\frac{\beta+\alpha}{2}\right) \leq 2/h_i \\
&\Leftrightarrow \quad -h_i\alpha \leq 4 - 4\left(\frac{\beta+\alpha}{2}\right) h_i \text{ and } \frac{\beta}{2} \leq 2/h_i - \frac{\alpha}{2} \\
&\Leftrightarrow \quad \beta \leq 2/h_i - \alpha/2 \text{ and } \beta \leq 4/h_i - \alpha.
\end{aligned}$$

Figure 9 (where we plot all the constraints we have so far) enables to conclude that the discriminant Δ_i is strictly positive and the algorithm is stable when the following three

conditions are satisfied:

$$\begin{aligned} \alpha &\geq 0 \\ \alpha + 2\beta &\leq 4/h_i \\ \alpha + \beta &\geq 2\sqrt{\alpha/h_1}. \end{aligned}$$

For any of those α et β we will have:

$$\eta_n^i = c_1(r_i^-)^n + c_2(r_i^+)^n.$$

Since $\eta_0^i = 0$, $c_1 + c_2 = 0$ and for $n = 1$, $c_1 = \eta_1^i/(r_i^- - r_i^+)$; we thus have:

$$\eta_n^i = \frac{\eta_1^i}{2} \frac{(r_i^+)^n - (r_i^-)^n}{\sqrt{\Delta_i}}.$$

Thus, we get the final expression:

$$(\phi_n^i)^2 h_i = \frac{(\phi_1^i)^2}{4n^2} \frac{\left\{\left[r_i + \sqrt{\Delta_i}\right]^n - \left[r_i - \sqrt{\Delta_i}\right]^n\right\}^2}{\Delta_i/h_i}.$$

C.2 Proof of Lemma 2

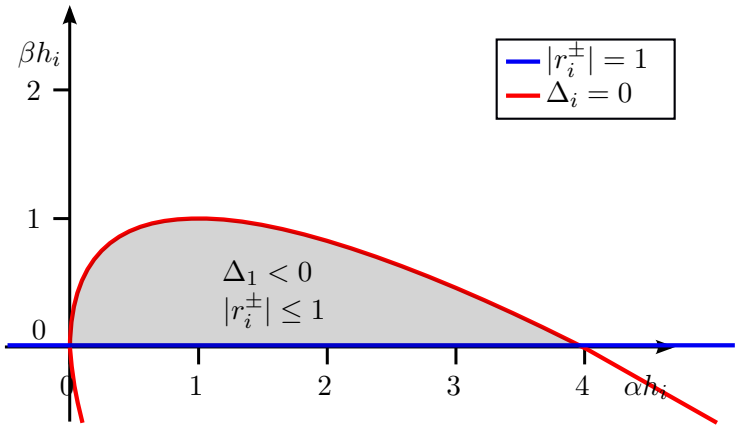

Figure 10: Stability in the complex case, with all constraints plotted.

The discriminant Δ_i is strictly negative if and only if $\left(\frac{\alpha+\beta}{2}\right)^2 h_i - \alpha < 0$. This implies $|\frac{\alpha+\beta}{2}| < \sqrt{\alpha/h_i}$. The modulus of the eigenvalues is $|r_i^{\pm}|^2 = 1 - \beta h_i$. Thus the discriminant Δ_i is strictly negative and the algorithm is stable for

$$\begin{aligned} \alpha, \beta &\geq 0 \\ \alpha + \beta &< \sqrt{\alpha/h_i}, \end{aligned}$$

as shown in Figure 10.

For any of those α et β we have:

$$\eta_n^i = [c_1 \cos(\omega_i n) + c_2 \sin(\omega_i n)]\rho_i^n,$$

with $\rho_i = \sqrt{1-\beta h_i}$, $\sin(\omega_i) = \sqrt{-\Delta_i}/\rho_i$ and $\cos(\omega_i) = r_i/\rho_i$. Since $\eta_0^i = 0$, $c_1 = 0$ and we have for $n = 1$, $c_2 = \eta_1^i/(\sin(\omega_i)\rho_i)$. Therefore

$$\eta_n^i = \eta_1^i \frac{\sin(\omega_i n)}{\sqrt{-\Delta_i}}(1-\beta h_i)^{n/2},$$

and

$$(\phi_n^i)^2 h_i = \frac{(\phi_1^i)^2}{n^2}\frac{\sin^2(\omega_i n)}{\sin^2(\omega_i)/h_i}(1-\beta h_i)^{n-1}.$$

C.3 Coalescing eigenvalues

When $\beta = 2\sqrt{\alpha/h_i} - \alpha$, the discriminant Δ_i is equal to zero and we have a double real eigenvalue:

$$r_i = 1 - \sqrt{\alpha h_i}.$$

Thus the algorithm is stable for $\alpha < \frac{4}{h_i}$. For any of those α et β we have:

$$\eta_n^i = (c_1 + n c_2) r^n.$$

This gives with $\eta_0^i = 0$, $c_1 = 0$ and $c_2 = \eta_1^i/r$. Therefore

$$\eta_n^i = n\eta_i(1-\sqrt{\alpha h_i})^{n-1},$$

and:

$$(\phi_n^i)^2 h_i = h_i(\phi_1^i)^2(1-\sqrt{\alpha h_i})^{2(n-1)}.$$

In the presence of coalescing eigenvalues the convergence is linear if $0 < \alpha < 4/h_i$ and $h_i > 0$, however one might worry about the behavior of $((\phi_n^i)^2 h_i)_n$ when h_i becomes small. Using the bound $x^2\exp(-x) \le 1$ for $x \le 1$, we have for $\alpha < 4/h_i$:

$$\begin{aligned} h_i(1-\sqrt{\alpha h_i})^{2n} &= h_i \exp(2n\log(|1-\sqrt{\alpha h_i}|)) \\ &\le hi\exp(-2n\min\{\sqrt{\alpha h_i}, 2-\sqrt{\alpha h_i}\}) \\ &\le \frac{h_i}{\min\{\sqrt{\alpha h_i}, 2-\sqrt{\alpha h_i}\}^2} \\ &\le \max\left\{\frac{1}{\alpha}, \frac{h_i}{(2-\sqrt{\alpha h_i})^2}\right\}. \end{aligned}$$

Therefore we always have the following bound for $\alpha < 4/h_i$:

$$(\phi_n^i)^2 h_i \le \frac{(\phi_1^i)^2}{4n^2}\max\left\{\frac{1}{\alpha}, \frac{h_i}{(2-\sqrt{\alpha h_i})^2}\right\}.$$

Thus for $\alpha h_i \le 1$ we get:

$$(\phi_n^i)^2 h_i \le \frac{(\phi_1^i)^2}{4n^2\alpha}.$$

D Proof of Theorem 2

D.1 Sketch of the proof

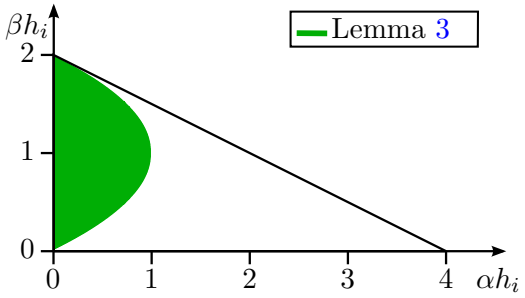

Figure 11: Validity of Lemma 3

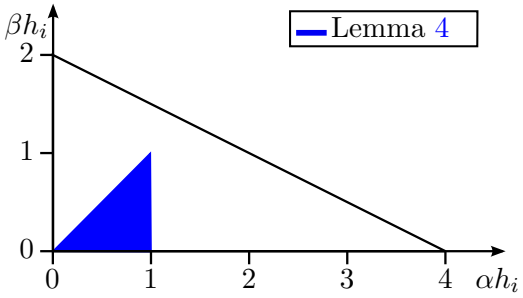

Figure 12: Validity of Lemma 4

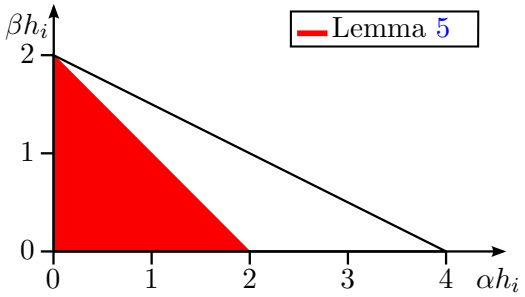

Figure 13: Validity of Lemma 5

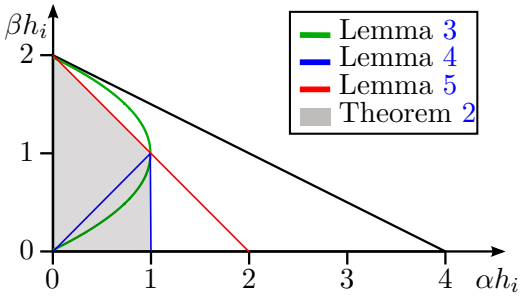

Figure 14: Area of Theorem 2

We divide the domain of validity of Theorem 2 in three subdomains as explained in Figure 14. On the domain described in Figure 11 we have a first bound on the iterate η_n^i:

Lemma 3. *For $0 \leq \alpha \leq 1/h_i$ and $1 - \sqrt{1-\alpha h_i} < \beta h_i < 1 + \sqrt{1-\alpha h_i}$, we have:*

$$(\eta_n^i)^2 \leq \frac{(\eta_1^i)^2}{\alpha h_i}.$$

And on the domain described Figure 12 we also have:

Lemma 4. *For $0 \leq \alpha \leq 1/h_i$ and $\beta \leq \alpha$ we have:*

$$(\eta_n^i)^2 \leq \frac{2(\eta_1^i)^2}{\alpha h_i}.$$

These two lemmas enable us to prove the first bound of Theorem 2 since the domain of this theorem is included in the intersection of the two domains of these lemmas as shown in Figure 14.

Then we have the following bound on domain described in Figure 13:

Lemma 5. *For $0 \leq \alpha \leq 2/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$, we have:*

$$|\eta_n^i| \leq \min\left\{\frac{2\sqrt{2n}}{\sqrt{(\alpha+\beta)h_i}}, \frac{4}{(\alpha+\beta)h_i}\right\}.$$

Since the domain of definition of Theorem 2 is included in the domain of definition of Lemma 5 (as shown in Figure 14), this lemma proves the last two bounds of the theorem.

D.2 Outline of the proofs of the Lemmas

- We find a Lyapunov function G from $\mathbb{R}^2$ to $\mathbb{R}$ such that the sequence $(G(\eta_n^i, \eta_{n-1}^i))$ decrease along the iterates.
- We also prove that $G(\eta_n^i, \eta_{n-1}^i)$ dominates $c\|\eta_n^i\|^2$ when we want to have a bound on $\|\eta_n^i\|^2$ of the form $\frac{1}{c}G(\eta_1^i, \eta_0^i) = \frac{1}{c}G(\theta_0^i - \theta_*^i, 0)$.

For readability, we remove the index i and take $h_i = 1$ without loss of generality.

D.3 Proof of Lemma 3

We first consider a quadratic Lyapunov function $\begin{pmatrix}\eta_n \\ \eta_{n-1}\end{pmatrix}^\top G_1 \begin{pmatrix}\eta_n \\ \eta_{n-1}\end{pmatrix}$ with $G_1 = \begin{pmatrix}1 & \alpha-1 \\ \alpha-1 & 1-\alpha\end{pmatrix}$.
We note that G_1 is symmetric positive semi-definite for $\alpha \leq 1$. We recall $F_i = \begin{pmatrix}2-(\alpha+\beta) & \beta-1 \\ 1 & 0\end{pmatrix}$.

For the result to be true we need for $0 \leq \alpha \leq 1$ and $1 - \sqrt{1-\alpha} < \beta < 1 + \sqrt{1-\alpha}$ two properties:

$$F_i^\top G_1 F_i \preccurlyeq G_1, \tag{19}$$

and

$$\alpha \begin{pmatrix}1 & 0 \\ 0 & 0\end{pmatrix} \preccurlyeq G_1. \tag{20}$$

Proof of Eq. (20). We have:

$$G_1 - \alpha \begin{pmatrix}1 & 0 \\ 0 & 0\end{pmatrix} = (1-\alpha)\begin{pmatrix}1 & 1 \\ 1 & 1\end{pmatrix} \succcurlyeq 0 \quad \text{for } \alpha \leq 1.$$

Proof of Eq. (19). Since $\beta \mapsto F_i(\beta)^\top G_1 F_i(\beta) - G_1$ is convex in β (G_1 is symmetric positive semi-definite), we only have to show Eq. (19) for the boundaries of the interval in β. For $x \in \mathbb{R}_+^*$:

$$\begin{pmatrix}x^2-x & x \\ 1 & 0\end{pmatrix}^\top \begin{pmatrix}1 & -x^2 \\ -x^2 & x^2\end{pmatrix} \begin{pmatrix}x^2-x & x \\ 1 & 0\end{pmatrix} - \begin{pmatrix}1 & -x^2 \\ -x^2 & x^2\end{pmatrix} = -(1-x^2)^2 \begin{pmatrix}1 & 0 \\ 0 & 0\end{pmatrix} \preccurlyeq 0.$$

This especially shows Eq. (19) for the boundaries of the interval with $x = \pm\sqrt{1-\alpha}$.

Bound. Thus, because $\eta_0 = 0$, we have

$$\alpha\eta_{n+1}^2 \leq \Theta_n^\top G_1 \Theta_n \leq \Theta_{n-1}^\top G_1 \Theta_{n-1} \leq \Theta_0^\top G_1 \Theta_0 \leq \eta_1^2.$$

This shows that for $0 \leq \alpha \leq 1/h_i$ and $1 - \sqrt{1-\alpha h_i} < \beta h_i < 1 + \sqrt{1-\alpha h_i}$:

$$(\eta_n^i)^2 \leq \frac{(\eta_1^i)^2}{\alpha h_i}.$$

D.4 Proof of Lemma 4

We consider now a second Lyapunov function $G_2(\eta_n, \eta_{n-1}) = (\eta_n - r\eta_{n-1})^2 - \Delta(\eta_{n-1})^2$. We have:

$$\begin{aligned}
G_2(\eta_n, \eta_{n-1}) &= (\eta_n - r\eta_{n-1})^2 - \Delta\eta_{n-1}^2 \\
&= (r\eta_{n-1} - (1-\beta)\eta_{n-2})^2 - \Delta\eta_{n-1}^2 \\
&= (r^2 - \Delta)\eta_{n-1}^2 + (1-\beta)^2\eta_{n-2}^2 - 2(1-\beta)r\eta_{n-1}\eta_{n-2} \\
&= ((1-\beta)\eta_{n-1}^2 + (1-\beta)(r^2-\Delta)\eta_{n-2}^2 - 2(1-\beta)r\eta_{n-1}\eta_{n-2} \\
&= (1-\beta)[(\eta_{n-1} - r\eta_{n-2})^2 - \Delta(\eta_{n-2})^2]. \\
&= (1-\beta)G_2(\eta_{n-1}, \eta_{n-2}).
\end{aligned}$$

Where we have used twice $r^2 - \Delta = (1-\beta)$ and $\eta_n = 2r\eta_{n-1} - (1-\beta)\eta_{n-2}$. Moreover $G_2(\eta_n, \eta_{n-1})$ can be rewritten as:

$$G_2(\eta_n, \eta_{n-1}) = (1 - \frac{\alpha+\beta}{2})(\eta_n - \eta_{n-1})^2 + \frac{\alpha-\beta}{2}(\eta_{n-1})^2 + \frac{\alpha+\beta}{2}(\eta_n)^2.$$

Thus for $\alpha + \beta \leq 2$ and $\beta \leq \alpha$ we have:

$$\frac{\alpha}{2}(\eta_n)^2 \leq G_2(\eta_n, \eta_{n-1}) = (1-\beta)^{n-1}G_2(\eta_1, \eta_0) = (1-\beta)^{n-1}(\eta_1)^2.$$

Therefore for $\alpha + \beta \leq 2/h_i$ and $\beta \leq \alpha$, we have:

$$(\eta_n^i)^2 \leq \frac{2(\eta_1^i)^2}{\alpha h_i}.$$

D.5 Proof of Lemma 5

We may write η_n as

$$\eta_n = r\eta_{n-1} + (r_+)^n + (r_-)^n.$$

Moreover, we have:

$$|(r_+)^n + (r_-)^n| \leq 2,$$

therefore for $\alpha + \beta \leq 2$,

$$|\eta_n| \leq r|\eta_{n-1}| + 2 \leq 2\frac{1-r^n}{1-r} \leq 2\frac{1-(1-(\frac{\alpha+\beta}{2}))^n}{(\frac{\alpha+\beta}{2})}.$$

Thus

$$|\eta_n| \leq \frac{2}{(\frac{\alpha+\beta}{2})h}.$$

Moreover for all $u \in [0,1]$ and $n \geq 1$ we have $1-(1-u)^n \leq \sqrt{nu}$, since $1-(1-u)^n \leq 1$ and $1-(1-u)^n = u\sum(1-u)^k \leq nu$. Thus

$$|\eta_n| \leq \frac{2\sqrt{n}}{\sqrt{(\frac{\alpha+\beta}{2})}}.$$

Therefore for $0 \leq \alpha \leq 2/h_i$ and $\alpha+\beta \leq 2/h_i$ we have:

$$|\eta_n^i| \leq \min\left\{\frac{2\sqrt{2n}}{\sqrt{(\alpha+\beta)h_i}}, \frac{4}{(\alpha+\beta)h_i}\right\}.$$

E Lower bound

We have the following lower-bound for the bound shown in Corollary 1, which shows that depending on which of the two terms dominates, we may always find a sequence of functions that makes it tight.

Proposition 1. *Let $L \geq 0$. For all sequences $0 \leq \alpha_n \leq 1/L$ and $0 \leq \beta_n \leq 2/L - \alpha_n$, such that $\alpha_n + \beta_n = o(n\alpha_n)$ there exists a sequence of one-dimensional quadratic functions $(f_n)_n$ with second-derivative less than L such that:*

$$\lim \alpha_n n^2 (f_n(\theta_n) - f_n(\theta_*)) = \frac{\|\theta_0 - \theta_*\|^2}{2}.$$

For all sequences $0 \leq \alpha_n \leq 1/L$ and $0 \leq \beta_n \leq 2/L - \alpha_n$, such that $n\alpha_n = o(\alpha_n + \beta_n)$, there exists a sequence of one-dimensional quadratic functions $(g_n)_n$ with second-derivative less than L such that:

$$\lim n(\alpha_n + \beta_n)(g_n(\theta_n) - g_n(\theta_*)) = \frac{(1-\exp(-2))^2 \|\theta_0 - \theta_*\|^2}{4}.$$

Proof of the first lower-bound. For the first lower bound we consider $0 \leq \alpha_n \leq 1/L$ and $0 \leq \beta_n \leq 2/L - \alpha$, such that $\alpha_n + \beta_n = o(n\alpha_n)$. We define $f_n = \pi^2/(4\alpha_n n^2)$ and we consider the sequence of quadratic functions $f_n(\theta) = \frac{f_n \theta^2}{2}$. We consider the iterate $(\eta_n)_n$ defined by our algorithm. We will show that

$$\lim \alpha_n f_n(\eta_n) = \frac{\eta_1^2}{2}.$$

We have, from Lemma 2,

$$f_n(\eta_n) = \frac{\eta_n^2 f_n}{2} = \frac{\eta_1^2 \sin^2(\omega_n n)\rho_n^{2n}}{2\alpha_n(1 - \frac{\pi^2(\alpha_n+\beta_n)^2}{(4\alpha_n n)^2})}.$$

Moreover,

$$\rho_n^{2n} = \left(1 - \frac{\beta_n \pi^2}{4\alpha_n n^2}\right)^n = \exp\left(n \log\left(1 - \frac{\beta_n \pi^2}{4\alpha_n n^2}\right)\right) = 1 + o(1),$$

since $\frac{\beta_n}{\alpha_n n} = o(1)$. Also, $1 - \frac{\pi^2(\alpha_n+\beta_n)^2}{(4\alpha_n n)^2} = 1 + o(1)$, since $\alpha_n + \beta_n = o(n\alpha_n)$. Moreover

$$\sin(\omega_n) = \frac{\sqrt{-\Delta_n}}{\rho_n} = \frac{\sqrt{f_n}\sqrt{\alpha_n - \frac{(\alpha_n+\beta_n)^2}{4} f_n}}{\sqrt{1-\beta_n f_n}} = \pi/(2n) + o(1/n),$$

thus $\omega_n = \pi/(2n) + o(1/n)$ and $\sin(n\omega_n) = 1 + o(1)$.

Proof of the second lower-bound. We consider now the situation where the second bound is active. Thus we take sequences (α_n) and (β_n), such that $n\alpha_n = o(\alpha_n + \beta_n)$. We define $g_n = \frac{2}{n(\alpha_n+\beta_n)} + \frac{4\alpha_n}{(\alpha_n+\beta_n)^2}$ and consider the sequence of quadratic functions $g_n(\theta) = \frac{g_n \theta^2}{2}$. We will show for the iterate (η_n) defined by our algorithm that:

$$\lim n(\alpha_n + \beta_n)(g_n(\theta_n) - g_n(\theta_*)) = \frac{(1 - \exp(-2))^2 \left\|\theta_0 - \theta_*\right\|^2}{4}.$$

We will use Lemma 1. We first have

$$\Delta_n = \left(\frac{\alpha_n + \beta_n}{2}\right)^2 g_n^2 - \alpha_n g_n = g_n \left(\frac{\alpha_n + \beta_n}{2}\right) \frac{1}{n}.$$

Thus $(n\Delta_n)/g_n = \left(\frac{\alpha_n+\beta_n}{2}\right)$ and

$$\begin{aligned}
\sqrt{\Delta_n} &= \sqrt{\left(\frac{1}{n}\right)^2 + \frac{2\alpha_n}{n(\alpha_n+\beta_n)}} \\
&= \frac{1}{n}\sqrt{1 + \frac{2\alpha_n n}{\alpha_n + \beta_n}} \\
&= \frac{1}{n} + \frac{\alpha_n}{\alpha_n+\beta_n} + o\left(\frac{\alpha_n}{\alpha_n+\beta_n}\right).
\end{aligned}$$

Moreover

$$r_n = 1 - \frac{\alpha_n + \beta_n}{2} g_n = 1 - \frac{1}{n} - \frac{2\alpha_n}{\alpha_n + \beta_n}.$$

Thus

$$r_+ = 1 - \frac{\alpha_n}{\alpha_n + \beta_n} + o\left(\frac{\alpha_n}{\alpha_n+\beta_n}\right),$$

and

$$r_+^n = \exp(n \log(r_+)) = \exp\left(-\frac{n\alpha_n}{\alpha_n+\beta_n}\right) + o\left(\frac{n\alpha_n}{\alpha_n+\beta_n}\right) = 1 + o(1).$$

Furthermore

$$r_- = 1 - \frac{2}{n} - \frac{3\alpha_n}{\alpha_n + \beta_n} + o\left(\frac{\alpha_n}{\alpha_n+\beta_n}\right),$$

and

$$r_-^n = \exp(n\log(r_+)) = \exp\left(-2 - \frac{3\alpha_n n}{\alpha_n+\beta_n}\right) + o\left(\frac{n\alpha_n}{\alpha_n+\beta_n}\right) = \exp(-2) + o(1).$$

Thus

$$(r_+^n - r_-^n)^2 = (1-\exp(-2))^2 + o(1).$$

Finally, we have:

$$\begin{aligned}(\alpha_n+\beta_n)n[g_n(\theta_n) - g_n(\theta_*)] &= \frac{\alpha_n+\beta_n}{2n}\|\theta_0-\theta_*\|^2\frac{[r_+^n - r_-^n]^2}{4\Delta_n/g_n} \\ &= \frac{\|\theta_0-\theta_*\|^2}{4}[r_+^n - r_-^n]^2 \\ &= \frac{\|\theta_0-\theta_*\|^2}{4}(1-\exp(-2))^2 + o(1).\end{aligned}$$

F Proofs of Section 4

F.1 Proofs of Theorem 3 and Theorem 4

We decompose again vectors in an eigenvector basis of H with $\eta_n^i = p_i^\top \eta_n$ and $\varepsilon_n^i = p_i^\top \varepsilon_n$:

$$\eta_{n+1}^i = (1-\alpha h_i)\eta_n^i + (1-\beta h_i)(\eta_n^i - \eta_{n-1}^i) + (n\alpha+\beta)\varepsilon_{n+1}^i.$$

We denote by $\xi_{n+1}^i = \begin{pmatrix}[n\alpha+\beta]\varepsilon_{n+1}^i \\ 0\end{pmatrix}$ and we have the reduced equation:

$$\Theta_{n+1}^i = F_i\Theta_n^i + \xi_{n+1}^i.$$

Unfortunately F_i is not Hermitian and this formulation will not be convenient for calculus. Without loss of generality, we assume $r_i^- \neq r_i^+$ even if it means having $r_i^- - r_i^+$ goes to 0 in the final bound. Let $Q_i = \begin{pmatrix} r_i^- & r_i^+ \\ 1 & 1\end{pmatrix}$ be the transfer matrix of F_i, i.e., $F_i = Q_i D_i Q_i^{-1}$ with $D_i = \begin{pmatrix} r_i^- & 0 \\ 0 & r_i^+\end{pmatrix}$ and $Q_i^{-1} = \frac{1}{r_i^- - r_i^+}\begin{pmatrix} 1 & -r_i^+ \\ -1 & r_i^-\end{pmatrix}$. We can reparametrize the problem in the following way:

$$\begin{aligned}Q_i^{-1}\Theta_{n+1}^i &= Q_i^{-1}F_i\Theta_n^i + Q_i^{-1}\xi_{n+1}^i \\ &= Q_i^{-1}F_iQ_iQ_i^{-1}\Theta_n^i + Q_i^{-1}\xi_{n+1}^i \\ &= D_i(Q_i^{-1}\Theta_n^i) + Q_i^{-1}\xi_{n+1}^i.\end{aligned}$$

With $\tilde{\Theta}_n^i = Q_i^{-1}\Theta_n^i$ and $\tilde{\xi}_n^i = Q_i^{-1}\xi_n^i$ we now have:

$$\tilde{\Theta}_{n+1}^i = D_i\tilde{\Theta}_n^i + \tilde{\xi}_{n+1}^i, \tag{21}$$

with now D_i Hermitian (even diagonal).

Thus it is easier to tackle using standard techniques for stochastic approximation (see, e.g., Polyak and Juditsky, 1992; Bach and Moulines, 2011):

$$\tilde{\Theta}_n^i = D_i^n \tilde{\Theta}_0^i + \sum_{k=1}^{n} D_i^{n-k} \tilde{\xi}_k^i.$$

Let $M_i = \begin{pmatrix} h_i^{1/2} & h_i^{1/2} \\ 0 & 0 \end{pmatrix}$, we then get using standard martingale square moment inequalities, since for $n \neq m$, ε_n^i and ε_m^i are uncorrelated (i.e., $\mathbb{E}[\varepsilon_n^i \varepsilon_m^i] = 0$):

$$\mathbb{E}\|M_i \tilde{\Theta}_n^i\|^2 = \|M_i D_i^n \tilde{\Theta}_0^i\|^2 + \mathbb{E} \sum_{k=1}^{n} \|M_i D_i^{n-k} \tilde{\xi}_k^i\|^2.$$

This is a bias-variance decomposition; the left term only depends on the initial condition and the right term only depends on the noise process.

We have with $M_i = \begin{pmatrix} h_i^{1/2} & h_i^{1/2} \\ 0 & 0 \end{pmatrix}$, $M_i Q_i^{-1} = \begin{pmatrix} 0 & h_i^{1/2} \\ 0 & 0 \end{pmatrix}$, and $M_i \tilde{\Theta}_n^i = \begin{pmatrix} \sqrt{h_i}\eta_n^i \\ 0 \end{pmatrix}$. Thus, we have access to the function values through:

$$\|M_i \tilde{\Theta}_n^i\|^2 = h_i (\eta_n^i)^2.$$

Moreover we have $\Theta_0^i = \begin{pmatrix} \phi_1^i/(r_i^- - r_i^+) \\ -\phi_1^i/(r_i^- - r_i^+) \end{pmatrix}$. Thus

$$\|M_i D_i^n \tilde{\Theta}_0^i\|^2 = (\phi_1^i)^2 h_i \frac{\left((r_i^+)^n - (r_i^-)^n\right)^2}{(r_i^+ - r_i^-)^2}.$$

This is the bias term we have studied in Section 3.3 which we bound with Theorem 2. The variance term is controlled by the next proposition.

Proposition 2. *With $\mathbb{E}[(\varepsilon_n^i)^2] = c_i$ for all $n \in \mathbb{N}$, for $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$, we have*

$$\frac{1}{n^2} \mathbb{E} \sum_{k=1}^{n} \|M_i D_i^{n-k} \tilde{\xi}_k^i\|^2 \leq \min\left\{ \frac{2(\alpha n + \beta)^2}{\alpha\beta(4 - (\alpha + 2\beta)h_i) n^2} \frac{c_i}{h_i}, \frac{16(n\alpha + \beta)^2}{n(\alpha + \beta)^2} \frac{c_i}{h_i}, 2\frac{(\alpha n + \beta)^2}{n\alpha} c_i, \frac{8(n\alpha + \beta)^2}{\alpha + \beta} c_i \right\}.$$

The last two bounds prove Theorem 3.

We note that if we restrict β to $\beta \leq 3/(2h_i) - \alpha/2$, then $4 - (\alpha + 2\beta)h_i \geq 1$ and the first bound of Proposition 2 is simplified to $\frac{2(\alpha n + \beta)^2}{\alpha\beta n^2} \frac{c_i}{h_i}$. This allows to conclude to prove Theorem 4.

F.2 Proof of Corollary 3

We let $\nu = \frac{\|\theta_0 - \theta_*\|}{\sqrt{L \operatorname{tr}(CH^{-1})}}$ and consider three different regimes depending on ν and L.

If $\nu < 1/L$, we have $\nu/N < 1/L$ and thus $\alpha = \nu/N$ and $\beta = \nu$. Therefore

$$\begin{aligned}\frac{\|\theta_0 - \theta_*\|^2}{N^2\alpha} + \frac{(\alpha N + \beta)^2}{\alpha\beta N^2}\operatorname{tr}(CH^{-1}) &= \frac{\|\theta_0 - \theta_*\|^2}{\nu N} + \frac{4\operatorname{tr}(CH^{-1})}{N} \\ &\leq \frac{\sqrt{L\operatorname{tr}(CH^{-1})}\|\theta_0 - \theta_*\|}{N} + \frac{4\operatorname{tr}(CH^{-1})}{N} \\ &\leq \frac{5\operatorname{tr}(CH^{-1})}{N},\end{aligned}$$

where we have used $\sqrt{L}\|\theta_0 - \theta_*\| < \sqrt{\operatorname{tr}(CH^{-1})}$ since $\nu < 1/L$.

If $\nu > 1/L$ and $\nu < N/L$, we have $\alpha = \nu/N$ and $\beta = 1/L$. Therefore

$$\begin{aligned}\frac{\|\theta_0 - \theta_*\|^2}{N^2\alpha} + \frac{(\alpha N + \beta)^2}{\alpha\beta N^2}\operatorname{tr}(CH^{-1}) &\leq \frac{\|\theta_0 - \theta_*\|^2}{\nu N} + \frac{4\operatorname{tr}(CH^{-1})}{L\nu N} \\ &\leq \frac{\sqrt{L\operatorname{tr}(CH^{-1})}\|\theta_0 - \theta_*\|}{N} + \frac{4\operatorname{tr}(CH^{-1})}{N} \\ &\leq \frac{5\sqrt{L\operatorname{tr}(CH^{-1})}\|\theta_0 - \theta_*\|}{N},\end{aligned}$$

where we have used $\sqrt{L}\|\theta_0 - \theta_*\| > \sqrt{\operatorname{tr}(CH^{-1})}$ since $\nu > 1/L$.

If $\nu > N/L$, we have $\alpha = 1/L$ and $\beta = 1/L$. Therefore

$$\begin{aligned}\frac{\|\theta_0 - \theta_*\|^2}{N^2\alpha} + \frac{(\alpha(N-1) + \beta)^2}{\alpha\beta N^2}\operatorname{tr}(CH^{-1}) &= \frac{L\|\theta_0 - \theta_*\|^2}{N^2} + \operatorname{tr}(CH^{-1}) \\ &\leq \frac{L\|\theta_0 - \theta_*\|^2}{N^2} + \frac{L\|\theta_0 - \theta_*\|^2}{N^2} \\ &\leq \frac{2L\|\theta_0 - \theta_*\|^2}{N^2},\end{aligned}$$

where we have used that the real bound in Proposition 2 is in fact in $(N-1)\alpha + \beta$, (see Lemma 6) and that $\operatorname{tr}(CH^{-1}) < \frac{L\|\theta_0 - \theta_*\|^2}{N^2}$ since $\nu > N/L$.

F.3 Proof of Proposition 2

F.3.1 Proof outline

To prove Proposition 2 we will use Lemmas 6, 7 and 8, that are stated and proved in Section F.3.2.

We want to bound $\mathbb{E}[\sum_{k=1}^n \|M_i D_i^{n-k} \tilde{\xi}_k^i\|^2]$ and according to Lemma 6, we have an explicit expansion using the roots of the characteristic polynomial:

$$\mathbb{E}\|M_i D_i^k \tilde{\xi}_k^i\|^2 = h_i((k-1)\alpha + \beta)^2 \mathbb{E}[(\varepsilon^i)^2] \frac{[(r_i^-)^{n-k} - (r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2}.$$

Thus, by bouding $(k-1)\alpha+\beta$ by $(n-1)\alpha+\beta$, we get

$$\mathbb{E}\sum_{k=1}^{n}\|M_iD_i^{n-k}\tilde{\xi}_k^i\|^2 \leq h_i((n-1)\alpha+\beta)^2\mathbb{E}[\varepsilon^{i2}]\sum_{k=1}^{n}\frac{[(r_i^-)^{n-k}-(r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2}. \tag{22}$$

Then, we have from Lemma 7 the inequality:

$$\sum_{k=1}^{n}\frac{[(r_i^-)^k-(r_i^+)^k]^2}{[(r_i^-)-(r_i^+)]^2} \leq \frac{2-\beta h_i}{4\alpha\beta h_i^2(1-(\frac{1}{4}\alpha+\frac{1}{2}\beta)h_i)}.$$

Therefore

$$\mathbb{E}\sum_{k=1}^{n}\|M_i^{1/2}D_i^{n-k}\tilde{\xi}_k^i\|^2 \leq \frac{\mathbb{E}[\varepsilon^{i2}]}{h_i}\frac{((n-1)\alpha+\beta)^2}{4\alpha\beta}\frac{2-\beta h_i}{(1-(\frac{1}{4}\alpha+\frac{1}{2}\beta)h_i)}.$$

This allows to prove the first part of the bound. The other parts are much simpler and are done in Lemma 8. Thus, adding these bounds gives for $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$:

$$\frac{1}{n^2}\mathbb{E}\sum_{k=1}^{n}\|M_iD_i^{n-k}\tilde{\xi}_k^i\|^2 \leq$$
$$\min\left\{\frac{2(\alpha(n-1)+\beta)^2}{\alpha\beta n^2(4-(\alpha+2\beta)h_i)}\frac{c}{h_i}, \frac{16((n-1)\alpha+\beta)^2}{n(\alpha+\beta)^2}\frac{c}{h_i}, 2\frac{(\alpha(n-1)+\beta)^2}{n\alpha}c_i, \frac{8((n-1)\alpha+\beta)^2}{\alpha+\beta}c_i\right\}.$$

F.3.2 Some technical Lemmas

We first compute an explicit expansion of the noise term as a function of the eigenvalues of the dynamical system.

Lemma 6. *For all $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$ we have*

$$\mathbb{E}\|M_iD_i^k\tilde{\xi}_k^i\|^2 = h_i((k-1)\alpha+\beta)^2\mathbb{E}[(\varepsilon^i)^2]\frac{[(r_i^-)^{n-k}-(r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2}.$$

Proof. We first turn the Euclidean norm into a trace, using that $\mathrm{tr}[AB] = \mathrm{tr}[BA]$ for two matrices A and B and that $\mathrm{tr}[x] = x$ for a real x.

$$\mathbb{E}\|M_iD_i^{n-k}\tilde{\xi}_k^i\|^2 = \mathrm{Tr}D_i^{n-k}M_i{}^\top M_iD_i^{n-k}\mathbb{E}[\tilde{\xi}_k^i(\tilde{\xi}_k^i)^\top], \tag{23}$$

This enables us to separate the noise term from the rest of the formula. Then we compute the latter from the definition of $\tilde{\xi}_k^i$ in Eq. (21) :

$$\mathbb{E}[\tilde{\xi}_k^i(\tilde{\xi}_k^i)^\top] = \frac{((k-1)\alpha+\beta)^2}{(r_i^- - r_i^+)^2}\mathbb{E}[(\varepsilon^i)^2]\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}.$$

And the first part of Eq. (23) is equal to:

$$D_i^{n-k}M_i{}^\top M_iD_i^{n-k} = h_i\begin{pmatrix}(r_i^-)^{2(n-k)} & (r_i^-)^{(n-k)}-(r_i^+)^{(n-k)}\\ (r_i^-)^{(n-k)}-(r_i^+)^{(n-k)} & (r_i^+)^{2(n-k)}\end{pmatrix},$$

because $D_i = \begin{pmatrix} r_i^- & 0 \\ 0 & r_i^+ \end{pmatrix}$ and $M_i = \begin{pmatrix} h_i^{1/2} & h_i^{1/2} \\ 0 & 0 \end{pmatrix}$. Therefore:

$$\mathbb{E}\|M_i D_i^{n-k} \tilde{\xi}_k^i\|^2 = h_i \frac{((k-1)\alpha+\beta)^2}{(r_i^- - r_i^+)^2} \mathbb{E}[\varepsilon^{i2}][(r_i^-)^{n-k} - (r_i^+)^{n-k}]^2.$$

□

In the following leamma, we bound a certain sum of powers of the roots.

Lemma 7. *For all $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$ we have*

$$\sum_{k=1}^{n} \frac{\left[(r_i^-)^k - (r_i^+)^k\right]^2}{\left[(r_i^-) - (r_i^+)\right]^2} \leq \frac{2 - \beta h_i}{4\alpha\beta h_i^2 (1 - (\frac{1}{4}\alpha + \frac{1}{2}\beta)h_i)}.$$

We first note that when the two roots become close, the denominator and the numerator will go to zero, which prevents from bounding the numerator easily. We also note that this bound is very tight since the difference between the two terms goes to zero when n goes to infinity.

Proof. We first expand the square of the difference of the powers of the roots and compute their sums.

$$\begin{aligned}
\sum_{k=1}^{n} \left[(r_i^-)^k - (r_i^+)^k\right]^2 &= \sum_{k=1}^{n} \left[{r_i^+}^{2k} + {r_i^-}^{2k} - 2(r_i^+ r_i^-)^k\right] \\
&= \frac{1 - {r_i^+}^{2n}}{1 - {r_i^+}^2} + \frac{1 - {r_i^-}^{2n}}{1 - {r_i^-}^2} - 2\frac{1 - (r_i^+ r_i^-)^n}{1 - (r_i^+ r_i^-)} \\
&= \frac{1}{1 - {r_i^+}^2} + \frac{1}{1 - {r_i^-}^2} - \frac{2}{1 - (r_i^+ r_i^-)} - \left[\frac{{r_i^+}^{2n}}{1 - {r_i^+}^2} + \frac{{r_i^-}^{2n}}{1 - {r_i^-}^2} - 2\frac{(r_i^+ r_i^-)^n}{1 - (r_i^+ r_i^-)}\right] \\
&= \frac{1}{1 - {r_i^+}^2} + \frac{1}{1 - {r_i^-}^2} - \frac{2}{1 - (r_i^+ r_i^-)} - I_n,
\end{aligned}$$

with $I_n = \left[\frac{{r_i^+}^{2n}}{1-{r_i^+}^2} + \frac{{r_i^-}^{2n}}{1-{r_i^-}^2} - 2\frac{(r_i^+ r_i^-)^n}{1-(r_i^+ r_i^-)}\right]$.

This sum is therefore equal to the sum of one term we will compute explicitly and one other term which will go to zero. We have for the first term:

$$\begin{aligned}
\frac{1}{1 - {r_i^+}^2} + \frac{1}{1 - {r_i^-}^2} - \frac{2}{1 - (r_i^+ r_i^-)} = &\frac{(1 - {r_i^-}^2)(1 - (r_i^+ r_i^-)) - (1 - {r_i^-}^2)(1 - {r_i^+}^2)}{(1 - {r_i^+}^2)(1 - {r_i^-}^2)(1 - (r_i^+ r_i^-))} \\
&+ \frac{(1 - {r_i^+}^2)(1 - (r_i^+ r_i^-)) - (1 - {r_i^-}^2)(1 - {r_i^+}^2)}{(1 - {r_i^+}^2)(1 - {r_i^-}^2)(1 - (r_i^+ r_i^-))},
\end{aligned}$$

with

$$\begin{aligned}(1-r_i^{-2})(1-(r_i^+r_i^-))-(1-r_i^{-2})(1-r_i^{+2}) &= (1-r_i^{-2})[(1-(r_i^+r_i^-))-(1-r_i^{+2})]\\ &= r_i^+(1-r_i^{-2})(r_i^+-r_i^-),\end{aligned}$$

and

$$(1-r_i^{+2})(1-(r_i^+r_i^-))-(1-r_i^{-2})(1-r_i^{+2}) = -r_i^-(1-r_i^{-2})(r_i^+-r_i^-),$$

and

$$\begin{aligned}r_i^+(1-r_i^{-2})(r_i^+-r_i^-)-r_i^-(1-r_i^{-2})(r_i^+-r_i^-) &= (r_i^+-r_i^-)[r_i^+(1-r_i^{-2})-r_i^-(1-r_i^{-2})]\\ &= (r_i^+-r_i^-)[r_i^+-r_i^-+r_i^+r_i^-(r_i^+-r_i^-)]\\ &= (r_i^+-r_i^-)^2[1+r_i^+r_i^-].\end{aligned}$$

Therefore the first term is equal to:

$$\frac{1}{1-r_i^{+2}}+\frac{1}{1-r_i^{-2}}-\frac{2}{1-(r_i^+r_i^-)} = \frac{(r_i^+-r_i^-)^2[1+r_i^+r_i^-]}{(1-r_i^{+2})(1-r_i^{-2})(1-(r_i^+r_i^-))},$$

and the sum can be expanded as:

$$\sum_{k=1}^{n}\frac{[(r_i^-)^k-(r_i^+)^k]^2}{[r_i^--r_i^+]^2} = \frac{[1+r_i^+r_i^-]}{(1-r_i^{+2})(1-r_i^{-2})(1-(r_i^+r_i^-))}-J_n,$$

with $J_n=\frac{I_n}{[(r_i^-)-(r_i^+)]^2}$.

Then we simplify the first term of this sum using the explicit values of the roots. We recall $r_i^\pm = r_i \pm \sqrt{\Delta_i} = 1-\frac{\alpha+\beta}{2}h_i \pm \sqrt{\left(\frac{\alpha+\beta}{2}\right)^2 h_i^2-\alpha h_i}$, therefore

$$\begin{aligned}r_i^+r_i^- &= r_i^2-\Delta_i^2\\ &= \left(1-\left(\frac{\alpha+\beta}{2}\right)h_i\right)^2-\left[\left(\frac{\alpha+\beta}{2}\right)h_i\right]^2+\alpha h_i\\ &= 1-\beta h_i,\end{aligned}$$

and

$$\begin{aligned}(1-r_i^{+2})(1-r_i^{-2}) &= [(1-r_i^-)(1-r_i^+)][(1+r_i^+)(1+r_i^+)]\\ &= [(1-r_i+\sqrt{\Delta_i})((1-r_i-\sqrt{\Delta_i})][(1+r_i+\sqrt{\Delta_i})((1+r_i-\sqrt{\Delta_i})]\\ &= [(1-r_i)^2-\Delta_i][(1+r_i)^2-\Delta_i][(1-r_i)^2-\Delta_i]\\ &= 4\alpha h_i\left(1-\left(\frac{1}{4}\alpha+\frac{1}{2}\beta\right)h_i\right).\end{aligned}$$

Thus

$$\sum_{k=1}^{n}\frac{[(r_i^-)^k-(r_i^+)^k]^2}{[(r_i^-)-(r_i^+)]^2} = \frac{2-\beta h_i}{4\alpha\beta h_i^2(1-(\frac{1}{4}\alpha+\frac{1}{2}\beta)h_i)}-J_n.$$

Even if J_n will be asymptotically small, we want a non-asymptotic bound, thus we will show that J_n is always positive.

In the real case $[(r_i^-) - (r_i^+)]^2 \geq 0$ and using $a^2 + b^2 \geq 2ab$, for all $(a, b) \in \mathbb{R}^2$, we have

$$\frac{{r_i^+}^{2n}}{1 - {r_i^+}^2} + \frac{{r_i^-}^{2n}}{1 - {r_i^-}^2} \geq 2\frac{(r_i^+ r_i^-)^n}{\sqrt{(1 - {r_i^+}^2)(1 - {r_i^-}^2)}},$$

and using ${r_i^+}^2 + {r_i^-}^2 \geq 2r_i^- r_i^-$ we have

$$\sqrt{(1 - {r_i^+}^2)(1 - {r_i^-}^2)} \leq 1 - (r_i^+ r_i^-),$$

since

$$\begin{aligned}(1 - {r_i^+}^2)(1 - {r_i^-}^2) - [1 - (r_i^+ r_i^-)]^2 &= 1 - {r_i^+}^2 - {r_i^-}^2 + (r_i^+ r_i^-)^2 - 1 + 2r_i^+ r_i^- - (r_i^+ r_i^-)^2 \\ &= 2r_i^+ r_i^- - {r_i^+}^2 - {r_i^-}^2 \\ &\leq 0.\end{aligned}$$

Thus

$$\frac{{r_i^+}^{2n}}{1 - {r_i^+}^2} + \frac{{r_i^-}^{2n}}{1 - {r_i^-}^2} - 2\frac{(r_i^+ r_i^-)^n}{1 - (r_i^+ r_i^-)} \geq 0.$$

and $J_n \geq 0$ in the real case.

In the complex case, $[(r_i^-) - (r_i^+)]^2 \leq 0$, and using $z^2 + \bar{z}^2 \leq 2z\bar{z}$ for all $z \in \mathbb{C}$, we have

$$\frac{{r_i^+}^{2n}}{1 - {r_i^+}^2} + \frac{{r_i^-}^{2n}}{1 - {r_i^-}^2} \leq 2\frac{(r_i^+ r_i^-)^n}{\sqrt{(1 - {r_i^+}^2)(1 - {r_i^-}^2)}},$$

and using ${r_i^+}^2 + {r_i^-}^2 \leq 2r_i^- r_i^-$ we have

$$\sqrt{(1 - {r_i^+}^2)(1 - {r_i^-}^2)} \geq 1 - (r_i^+ r_i^-).$$

Thus

$$\frac{{r_i^+}^{2n}}{1 - {r_i^+}^2} + \frac{{r_i^-}^{2n}}{1 - {r_i^-}^2} - 2\frac{(r_i^+ r_i^-)^n}{1 - (r_i^+ r_i^-)} \leq 0.$$

and $J_n \geq 0$ in the complexe case.

Therefore we always have:

$$J_n \geq 0,$$

and

$$\sum_{k=1}^{n} \frac{[(r_i^-)^k - (r_i^+)^k]^2}{[(r_i^-) - (r_i^+)]^2} \leq \frac{2 - \beta h_i}{4\alpha\beta h_i^2 (1 - (\frac{1}{4}\alpha + \frac{1}{2}\beta)h_i)}.$$

□

However we can also bound roughly Eq. (22) using Theorem 2 since we recall we have $\eta_n^i = \frac{[(r_i^-)^{n-k} - (r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2}$. This gives us the following lemma which enables to prove the second part of Proposition 2.

Lemma 8. *For all $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$ we have*

$$\mathbb{E}\sum_{k=1}^n \|M_i^{1/2} D_i^{n-k} \tilde{\xi}_k^i\|^2 \leq \mathbb{E}[(\varepsilon^i)^2] n((n-1)\alpha + \beta)^2 \min\left\{\frac{2}{\alpha}, \frac{8n}{\alpha+\beta}, \frac{16}{h_i(\alpha+\beta)^2}\right\}.$$

Proof. From Lemma 6, we get

$$\begin{aligned}
\mathbb{E}\sum_{k=1}^n \|M_i^{1/2} D_i^{n-k} \tilde{\xi}_k^i\|^2 &= h_i \mathbb{E}[(\varepsilon^i)^2] \sum_{k=1}^n ((k-1)\alpha + \beta)^2 \frac{[(r_i^-)^{n-k} - (r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2} \\
&\leq h_i \mathbb{E}[(\varepsilon^i)^2] ((n-1)\alpha + \beta)^2 n \min\left\{\frac{2}{\alpha h_i}, \frac{8n}{(\alpha+\beta)h_i}, \frac{16}{(\alpha+\beta)^2 h_i^2}\right\} \\
&\leq \mathbb{E}[(\varepsilon^i)^2] n((n-1)\alpha + \beta)^2 \min\left\{\frac{2}{\alpha}, \frac{8n}{\alpha+\beta}, \frac{16}{h_i(\alpha+\beta)^2}\right\}.
\end{aligned}$$

□

G Comparison with additional other algorithms

G.1 Summary

When the objective function f is quadratic and for correct choices of step-sizes, the AC-SA algorithm of Lan (2012), the SAGE algorithm of Hu et al. (2009) and the Accelerated RDA algorithm of Xiao (2010) are all equivalent to:

$$\theta_{n+1} = [I - \delta_{n+1} H_{n+1}]\theta_n + \frac{n-2}{n+1}[I - \delta_{n+1} H_{n+1}](\theta_n - \theta_{n-1}) + \delta_{n+1}\varepsilon_{n+1},$$

where we use $H_n\theta + \varepsilon_n$ as an unbiased estimate of the gradient and δ_n as step-size which values will be specified later.

Lan (2012) and Hu et al. (2009) only consider bounded cases by projecting their iterates on a bounded space. Xiao (2010) deals with the unbounded case and prove the following convergence result:

Theorem 5. *(Xiao, 2010, Theorem 6). With $\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] = C$, for step-size $\delta_n \leq \frac{n-1}{n}\gamma$ with $\gamma \leq 1/L$, we have*

$$\mathbb{E}f(\theta_n) - f(\theta_*) \leq \frac{4\|\theta_0 - \theta_*\|^2}{n^2\gamma} + \frac{n\gamma\sigma^2 \operatorname{tr} C}{3}.$$

This result is significantly more general than ours since it is valid for composite optimization and general noise on the gradients.

We now present the different algorithms and show they all share the same form.

G.2 AC-SA

Lemma 9. *AC-SA algorithm with step size γ_n and β_n and gradient estimate $H_{n+1}\theta_n+\varepsilon_{n+1}$ is equivalent to:*

$$\theta_{n+1} = (I - \frac{\gamma_n}{\beta_n}H_{n+1})\theta_n + \frac{\beta_{n-1}-1}{\beta_n}(I - \frac{\gamma_n}{\beta_n}H_{n+1})(\theta_n - \theta_{n-1}) + \frac{\gamma_n}{\beta_n}\varepsilon_{n+1}.$$

Proof. We recall the general **AC-SA algorithm**:

- Let the initial points $x_1^{ag} = x_1$, and the step-sizes $\{\beta_n\}_{n\leq 1}$ and $\{\gamma_n\}_{n\leq 1}$ be given. Set $n = 1$
- **Step 1**. Set $x_n^{md} = \beta_n^{-1}x_n + (1-\beta_n^{-1})x_n^{ag}$,
- **Step 2**. Call the Oracle for computing $G(x_n^{md}, \xi_n)$ where $\mathbb{E}[G(x_n^{md}, \xi_n)] = f'(x_n^{md})$. Set
$$x_{n+1} = x_n - \gamma_n G(x_n^{md}, \xi_n),$$
$$x_{n+1}^{ag} = \beta_n^{-1}x_{n+1} + (1-\beta_n^{-1})x_n^{ag},$$
- **Step 3**. Set $n \to n+1$ and go to step 1.

When f is quadratic we will have $G(x_n^{md}, \xi_n) = H_{n+1}x_n^{md} - \varepsilon_{n+1}$, thus $x_{n+1} = x_n - \gamma_n H_{n+1}x_n^{md} + \gamma_n\varepsilon_{n+1}$, and:

$$\begin{aligned} x_{n+1}^{ag} &= \beta_n^{-1}x_{n+1} + (1-\beta_n^{-1})x_n^{ag} \\ &= \beta_n^{-1}(x_n - \gamma_n H_{n+1}x_n^{md} + \gamma_n\varepsilon_{n+1}) + (1-\beta_n^{-1})x_n^{ag} \\ &= \beta_n^{-1}(\beta_n x_n^{md} + (1-\beta_n)x_n^{ag} - \gamma_n H_{n+1}x_n^{md} + \gamma_n\varepsilon_{n+1}) + (1-\beta_n^{-1})x_n^{ag} \\ &= x_n^{md} - \frac{\gamma_n}{\beta_n}H_{n+1}x_n^{md} + \frac{\gamma_n}{\beta_n}\varepsilon_{n+1}, \end{aligned}$$

and

$$\begin{aligned} x_n^{md} &= \beta_n^{-1}x_n + (1-\beta_n^{-1})x_n^{ag} \\ &= \beta_n^{-1}\beta_{n-1}x_n^{ag} + \beta_n^{-1}(1-\beta_{n-1})x_{n-1}^{ag} + (1-\beta_n^{-1})x_n^{ag} \\ &= x_n^{ag} + \frac{\beta_{n-1}-1}{\beta_n}[x_n^{ag} - x_{n-1}^{ag}]. \end{aligned}$$

These give the result for $\theta_n = x_n^{ag}$. □

G.3 SAGE

Lemma 10. *The algorithm SAGE with step-sizes L_n and α_n is equivalent to:*

$$\theta_{n+1} = (I - 1/L_{n+1}H_{n+1})\theta_n + (1-\alpha_n)\frac{\alpha_{n+1}}{\alpha_n}[I - 1/L_{n+1}H_{n+1}](\theta_n - \theta_{n-1}) + 1/L_{n+1}\varepsilon_{n+1}.$$

Proof. We recall the general **SAGE algorithm**:

- Let the initial points $x_0 = z_0 = 0$, and the step-sizes $\{\beta_n\}_{n\leq 1}$ and $\{L_n\}_{n\leq 1}$ be given.
  Set $n = 1$
- **Step 1**. Set $x_n = (1-\alpha_n)y_{n-1} + \alpha_n z_{n-1}$,
- **Step 2**. Call the Oracle for computing $G(x_n, \xi_n)$ where $\mathbb{E}[G(x_n, \xi_n)] = f'(x_n)$. Set

$$y_n = x_n - 1/L_n G(x_n, \xi_n),$$

$$z_n = z_{n-1} - \alpha_n^{-1}(x_n - y_n)$$

- **Step 3**. Set $n \to n+1$ and go to step 1.

We have

$$y_n = (I - 1/L_n H_n)x_n + \gamma_n \varepsilon_n,$$

and

$$\begin{aligned} z_n &= z_{n-1} - \alpha_n^{-1}(x_n - y_n) \\ &= z_{n-1} - \alpha_n^{-1}[(1-\alpha_n)y_{n-1} + \alpha_n z_{n-1} - y_n] \\ &= \alpha_n^{-1} y_n - \alpha_n^{-1}(1-\alpha_n)y_{n-1}. \end{aligned}$$

Thus

$$\begin{aligned} x_n &= (1-\alpha_n)y_{n-1} + \alpha_n z_{n-1} \\ &= (1-\alpha_n)y_{n-1} + \alpha_n[\alpha_{n-1}^{-1} y_{n-1} - \alpha_{n-1}^{-1}(1-\alpha_{n-1})y_{n-2}] \\ &= y_{n-1} + (1-\alpha_{n-1})\frac{\alpha_n}{\alpha_{n-1}}[y_{n-1} - y_{n-2}]. \end{aligned}$$

These give the result for $\theta_n = y_n$. □

G.4 Accelerated RDA method

Lemma 11. *The algorithm AccRDA with step-sizes β and α_n is equivalent to:*

$$\theta_{n+1} = (I - \gamma_{n+1}H_{n+1})\theta_n + (1-\alpha_n)\frac{\alpha_{n+1}}{\alpha_n}[I - \gamma_{n+1}H_{n+1}](\theta_n - \theta_{n-1}) + \gamma_{n+1}\varepsilon_{n+1},$$

with $\gamma_n = \frac{\alpha_n \theta_n}{L+\beta}$.

Proof. We recall the general **Accelerated RDA method**:

- Let the initial points $w_0 = v_0$, $A_0 = 0$, $\tilde{g}_0 = 0$ and the step-sizes $\{\alpha_n\}_{n\leq 1}$ and $\{\beta_n\}_{n\leq 1}$ be given.
  Set $n = 1$

- **Step 1**. Set $A_n = A_{n-1} + \alpha_n$ and $\theta_n = \frac{\alpha_n}{A_n}$.
- **Step 2**. Compute the query point $u_n = (1-\theta_n)w_{n-1} + \theta_n v_{n-1}$
- **Step 3**. Call the Oracle for computing $g_n = G(u_n, \xi_n)$ where $\mathbb{E}[G(u_n, \xi_n)] = f'(u_n)$, and update the weighted average $\tilde{g}_n$

$$\tilde{g}_n = (1-\theta_n)\tilde{g}_{n-1} + \theta_n g_n.$$

- **Step 4**. Set $v_n = v_0 - \frac{A_n}{L+\beta_n}\tilde{g}_n$.
- **Step 5**. Set $w_n = (1-\theta_n)w_{n-1} + \theta_n v_n$.
- **Step 6**. Set $n \to n+1$ and go to step 1.

First we have

$$\begin{aligned}
v_n &= v_0 - \frac{A_n}{L+\beta_n}\tilde{g}_n \\
&= v_0 - \frac{A_n}{L+\beta_n}[(1-\theta_n)\tilde{g}_{n-1} + \theta_n g_n] \\
&= v_0 - \frac{A_n}{L+\beta_n}[(1-\theta_n)\tilde{g}_{n-1} + \theta_n(H_{n+1}u_n + \varepsilon_{n+1})] \\
&= v_0 + (1-\theta_n)\frac{A_n(L+\beta_{n-1})}{(L+\beta_n)A_{n-1}}v_{n-1} - \frac{A_n}{L+\beta_n}\theta_n(H_{n+1}u_n + \varepsilon_{n+1})] \\
&= v_0 + (1-\theta_n)\frac{A_n(L+\beta_{n-1})}{(L+\beta_n)A_{n-1}}v_{n-1} - \frac{\alpha_n}{L+\beta_n}(H_{n+1}u_n + \varepsilon_{n+1})].
\end{aligned}$$

With $\beta_n = \beta$ we have $v_n = v_{n-1} - \frac{\alpha_n}{L+\beta}(H_{n+1}u_n + \varepsilon_{n+1})]$ and

$$w_n = (I - \frac{\alpha_n\theta_n}{L+\beta}H_{n+1})u_n + \frac{\alpha_n\theta_n}{L+\beta}\varepsilon_{n+1}.$$

Since $v_{n-1} = \theta_{n-1}^{-1}w_{n-1} - \theta_{n-1}^{-1}(1-\theta_{n-1})w_{n-2}$, then

$$u_n = (1-\theta_n)w_{n-1} + \theta_n(\theta_{n-1}^{-1}w_{n-1} - \theta_{n-1}^{-1}(1-\theta_{n-1})w_{n-2}),$$

and

$$u_n = w_{n-1} + \frac{\alpha_n A_{n-2}}{\alpha_{n-1}A_n}[w_{n-1} - w_{n-2}].$$

□

H Lower bound for stochastic optimization for least-squares

In this section, we show a lower bound for optimization of quadratic functions with noisy access to gradients. We follow very closely the framework of Agarwal et al. (2012) and use

their notations. The only difference with their Theorem 1 in the different choice of two functions f_i^+ and f_i^-, which we choose to be:

$$f_i^{\pm}(x) = c_i(x_i \pm \frac{r}{2})^2,$$

with a non-increasing sequence (c_i) to be chosen later. The function g_α that is optimized is thus:

$$g_\alpha(x) = \frac{1}{d}\sum_{i=1}^{d} \big\{(\frac{1}{2} + \alpha_i\delta)f_i^+(x) + (\frac{1}{2} - \alpha_i\delta)f_i^-(x)\big\}.$$

This function is quadratic and its Hessian has eigenvalues equal to $2c_i/d$. Thus, its largest eigenvalue is $2c_1/d$, which we choose equal to L.

Noisy gradients are obtained by sampling d independent Bernoulli random variables b_i, $i = 1, \dots, d$, with parameters $(\frac{1}{2} + \alpha_i\delta)$ and using the gradient of the random function $\frac{1}{d}\sum_{i=1}^{d}\big\{b_i f_i^+(x) + (1-b_i)f_i^-(x)\big\}$. The variance of the random gradient is equal to

$$V = \sum_{i=1}^{d} \frac{1}{d^2}\mathrm{var}\Big(b_i\big[c_i(x_i + r/2) - c_i(x_i - r/2)\big]\Big) = \frac{1}{d^2}\sum_{i=1}^{d} c_i^2 r^2(1/4 - \delta^2).$$

The function g_α is minimized for $x = -\alpha\delta r$, and the discrepancy measure between two functions g_α and g_β is greater than

$$\frac{1}{d}\sum_{i=1}^{d}\left\{\inf_x\big\{f_i^+(x) + f_i^-(x)\big\} - \inf_x f_i^+(x) - \inf_x f_i^-(x)\right\}1_{\alpha_i\neq\beta_i} \geqslant \frac{1}{d}\sum_{i=1}^{d}\frac{3c_i r^2\delta^2}{4}1_{\alpha_i\neq\beta_i} \geqslant \frac{1}{d}\frac{3c_d r^2\delta^2}{4}\Delta(\alpha,\beta).$$

Since the vectors $\alpha, \beta \in \{-1,1\}^d$ are so that their Hamming distance $\Delta(\alpha,\beta) \geqslant d/4$ for $\alpha \neq \beta$, we have a discrepancy measure greater than $\frac{3c_d r^2\delta^2}{16}$. Thus, for a an approximate optimality of $\varepsilon = \frac{c_d r^2\delta^2}{38}$, we have, following the proof of Theorem 1 (equation (29)) from Agarwal et al. (2012), for N iterations of any method that accesses a random gradient, we have:

$$1/3 \geqslant 1 - 2\frac{16Nd\delta^2 + \log 2}{d\log(2/\sqrt{e})}.$$

Thus, for d large, we get, up to constants, $\delta^2 \geqslant 1/N$ and thus $\varepsilon \geqslant \frac{r^2 c_d}{N}$.

For $c_1 = 2Ld$ and $c_i = L\sqrt{d}$ for the remaining ones, we get (up to constants):

$$\varepsilon \geqslant \frac{V}{L}\frac{\sqrt{d}}{N}.$$

This leads to the desired result for $N \leqslant d$.